Modelle und Methoden für die Kopplung automatischer und visuell-interaktiver Verfahren für die Datenanalyse

Fachbereich Informatik (FB 20)
der Technischen Universität Darmstadt

Dissertation

zur Erlangung des akademischen Grades
eines Doktor–Ingenieurs (Dr.–Ing.)

von

Dipl.–Math. Thorsten May

geboren in Offenbach

Referent:	Prof. Dr.–techn. Dieter W. Fellner, Technische Universität Darmstadt
Korreferent:	Prof. Dr.–techn. Helwig Hauser, Universität Bergen
Tag der Einreichung:	23.9.2011
Tag der mündlichen Prüfung:	4.11.2011

Darmstädter Dissertationen
D 17
Darmstadt, 2012

Lasst mich Fragen stellen
nicht Dogmen in die Welt,
denn Fragen werden älter

Danksagung

Diese Arbeit ist das Ergebnis von mittlerweise sieben Jahren Forschung am Fraunhofer IGD in Darmstadt. Ein Kopf und zwei Hände allein haben vielleicht genügt, um diese Arbeit zu beginnen, aber nach dem so viel Zeit seit meinem ersten - und mir heute sehr fremd vorkommenden - Entwurf ins Land gezogen sind, kann ich sicher sagen, dass sie niemals genügt hätten, um diese Arbeit zu beenden. Für die Möglichkeit, in diesem jungen, zu einem guten Teil noch brachliegendem Forschungsgebiet ein paar Fußabdrücke hinterlassen zu dürfen, danke ich meinem Referenten und Betreuer Professor Dieter Fellner, der mir besonders geholfen hat, das Profil dieser Arbeit zu schärfen - trotz meiner Bestrebungen, am besten überall und umfassend und tief zu schürfen. Vielleicht wäre ich sonst immer noch nicht fertig. Ebenso danke ich meinem Koreferenten Professor Helwig Hauser, der den weiten Weg und die Arbeit auf sich genommen hat, um meine Promotion zu unterstützen und hilfreiche Vorschläge zur Verbesserung der Arbeit gemacht hat.
Weit mehr als nur meinen Broterwerb verdanke ich Jörn Kohlhammer. Visual Analytics war am Anfang für uns beide neu, und ohne sein Vertrauen und seine Beharrlichkeit wäre die Erschließung dieses Forschungsgebiets für unsere Abteilung und letztlich für diese Promotion schwer, wenn nicht gar unmöglich, geworden. Auch sein Talent, Leute aus unserem Forschungsgebiet zusammenzubringen, hat mir sehr geholfen, mich in der wissenschaftlichen Community heimisch zu fühlen.
Vor allen Dingen wird eine Dissertation besser durch (manchmal anstrengende) Fragen. Diese Fragen verdanke ich vor allem meinen Kollegen aus dem Institut, die das Talent haben, sie dann zu stellen, wenn mir zu einem Thema keine mehr in den Sinn kommen. Besonderen Dank gilt Sascha Schneider für viel Geduld in meinen ersten Jahren am Institut und Arnulph Fuhrmann, Jörg Sahm, Marcus Hoffmann, Brigitte Kötting, Martin Knuth und Tobias Ruppert, denen ich mein Thema so lange erklären musste, bis ich selbst wußte, um was es ging. Unendlich Hilfe erhielt ich auch von Gabriele Knöß, die verhindern konnte, dass ich irgendwo zwischen Reiseanträgen, Druckaufträgen und Umläufen den Faden verliere. Ich bin froh, dass ich mit Euch arbeiten durfte und darf.
Von der Forschungsgruppe GRIS danke ich besonders Tobias Schreck, Tatiana von Landesberger und Sebastian Bremm für zahlreiche anregende Diskussionen um eigene und fremde Publikationen. Arjan Kuijper danke ich dafür, dass er mit beharrlicher Regelmäßigkeit dafür sorgte, dass mir die Dissertation nicht aus dem Blickfeld gerät.
Ich danke auch meinen Studenten, besonders Dirk Mattheis, Andreas Bannach und Alexander Marinc, die mir geholfen haben, die Konzepte dieser Arbeit technisch umzusetzen, dafür, dass sie mir glaubten, dass das tatsächlich alles funktionieren könnte.

Ich habe auch meinen Freunden zu danken; einigen deshalb, weil ihre Kommentare mir bisweilen ein Stachel im Fleisch waren, diese Arbeit doch endlich zu beenden. Den anderen deshalb, weil sie mich nicht damit behelligt haben, so dass mir auch die Stunden fern des Rechners wirklich Erholung blieben. Sie alle haben mir mehr geholfen, als Ihnen vermutlich bewußt ist.
Der letzte Dank gilt meinen Eltern und meiner Schwester; sie haben mir mehr geholfen als *mir* vermutlich bewußt ist. Es gibt die schlechteren Tage, an denen man vergißt, wie ein Lächeln oder ein Danke geht, selbst wenn man weiß, dass man den Rücken freigehalten bekommt. Danke, dass Ihr da seid.

Thorsten May

Mühlheim, April 2012

Zusammenfassung

In dieser Arbeit werden neue Kopplungsvarianten von Visualisierungstechniken und Data-Mining-Verfahren für die Datenanalyse vorgestellt. Die betrachteten Teilschritte der Analyse sind hierbei die Suche von Mustern, deren Transformation in formale Modelle und die Überprüfung dieser Modelle. Die Suche nach Mustern und deren Modellierung ist in diesen Technologien bisher nicht getrennt. In diesem Konzept werden die Aufgaben neu aufgeteilt: Die Suche und Identifizierung von Mustern übernimmt der Mensch mit Hilfe von interaktiven Visualisierungstechniken. Die Transformation der gefundenen Muster wird durch Data-Mining Verfahren übernommen. Diese Aufteilung nutzt die spezifischen Stärken von Mensch und Maschine besser.

Der Mensch trägt mit seinem flexiblen und robusten Wahrnehmungssystem zu dieser Aufgabe bei; er wird gleichzeitig von der kognitiv anspruchsvolleren Aufgabe der Transformation der Muster entlastet. Die Maschine kann auch insbesondere komplexe Muster leichter in formale Modelle umformen. Wenn der Mensch die Interpretation ohne Hilfmittel oder Übung durchführen muß, können selten mehr als zwei bis dreidimensionale Muster interpretiert werden. In der Arbeit wird gezeigt, dass die Muster zehn- und mehrdimensionaler Zusammenhänge nicht nur dargestellt und wahrgenommen, sondern auch genutzt werden können.

Durch die Aufteilung kann auch jene Schwäche automatischer Verfahren kompensiert werden, dass sie nur solche Muster finden, für deren Suche sie entwickelt wurden. Bei der Konstruktion der Modelle eliminiert die interaktive Vorgabe Mehrdeutigkeiten in den Mustern. Zur Überprüfung des konstruierten Modells wird daraus wieder ein Muster generiert, das mit den Mustern der Originaldaten visuell verglichen wird. Dies liefert qualitative Information über die Art des Modellierungsfehlers und die prinzipielle Eignung des automatischen Verfahrens, die quantitative Gütemaße allein nicht bereitstellen können.

Zusammengenommen beschreibt das Konzept zwei Verbindungen zwischen visuell-interaktiven Techniken und automatischen Techniken in entgegengesetzte, sich ergänzende Richtungen. Mustererkennung, Interaktion und automatische Modellierung beschreiben den Weg von Muster zum Modell. Simulation, Feedback und visueller Abgleich beschreiben den Weg vom Modell zurück zum Muster. Das Modell der Daten wird in einem zyklischen, iterativen Prozess konstruiert, stetig überprüft und verfeinert.

Um diese beiden Varianten der Kopplung von anderen, existierenden Verfahren abgrenzen zu können, wurde das allgemeine Modell des Visual-Analytics-Prozess verfeinert. Unabhängig von spezifischen Techniken wurden acht verschiedene Kopplungsvarianten identifiziert. Dabei werden sowohl bekannte Ansätze in die Systematik eingeordnet, als auch solche, die bisher nicht oder kaum in der Literatur umgesetzt werden.

Januar 2011 Thorsten May

Inhaltsverzeichnis

Kapitel 1

Einleitung

„Data, don't babble.“ -
„Babble, sir ? I am not aware that I ever babble, sir. It may be that from time to time I have considerable information to communicate and you may question the way in which I organize it–“ -
„Please, organize it into brief answers to my questions, we have very little time. Do they accept our precence at this planet ?“ -
„Undecided, sir.“ -
„Data, please feel free to volunteer any important information.“
–Picard und Data, Star Trek: The Next Generation

Das letzte Ziel jeder Datenanalyse besteht darin, Entscheidungen zu verbessern. Die Komplexität unserer Welt kann man ermessen, wenn man die Reichweite und die Implikation der Entscheidungen eines Menschen vergleicht mit seinen Möglichkeiten, die Konsequenzen dieser Entscheidungen im Blick zu behalten. In manchen Fällen überspannen die Beziehungen zwischen Fakten und Zielen einen Zeitrahmen, der in Jahren oder sogar Jahrzehnten gemessen werden kann. In anderen Fällen sind die Beziehungen zu komplex, als dass sie ohne Hilfsmittel erkannt und in Entscheidungen umgesetzt werden können. In unserer so genannten schnelllebigen Zeit schiebt das Vordergrundrauschen des Alltagsgeschäfts die Reflexion und Verbesserung von langfristigen Entscheidungen ständig aus unserem Fokus. Dies gilt für sehr unterschiedliche Gebiete wie Finanzwirtschaft, Public-Policy Management, Ökologie, Biowissenschaften und zahlreiche mehr.
In Forschung und Wissenschaft ist mindestens eine methodische Schwäche, die Konsequenzen der eigenen Annahmen und Entscheidungen zu ignorieren. In der Finanzwirtschaft, Medizin oder Politik kann es unbezahlbar teuer wenn nicht sogar ethisch fragwürdig werden, diese Konsequenzen nicht abschätzen zu können.
Was sich im Großen auf die Fragestellungen eines ganzen Anwendungsgebietes bezieht, setzt sich im Kleinen in jeder Entscheidung in der Datenanalyse fort. Analyse, wie sie hier verstan-

den wird, ist kein „Frage-Antwort"-Prozess, denn dies würde bedeuten, dass der Mechanismus, der Antworten produziert, bereits bekannt ist. Aus der Perspektive des Analysten[1], ist die Analyse ein aktiver, konstruktiver und oft auch ein kreativer Prozess für die Entwicklung von Entscheidungsoptionen, Kriterien, Begriffen und Modellen. Als Konsequenz kann das Ergebnis einer Analyse ebenso von den Daten abhängig sein, wie vom Hintergrundwissen, der Erfahrung und den Annahmen des Analysten. Die Wahl einer Analysemethode, eines Modells oder selbst eines scheinbar unbedeutenden Parameters ist Teil dieses konstruktiven Prozesses.
Es gibt gute Gründe dafür, eine Analyse mit Annahmen zu beginnen, auf deren Basis Methoden und Modelle gewählt werden. Natürlicherweise basiert jeder Vorgang, in dem Wissen geschöpft oder geprüft werden soll, auf unvollständigem Wissen. Daraus folgt aber, dass jedes Ergebnis ebenso ein Artefakt der Daten, wie auch ein Artefakt der Annahmen des Analysten sein kann. Jede Annahme propagiert in der Analyse bis zum Ergebnis, und kann potentiell Entscheidungen und Konsequenzen beeinflussen. Ein Ziel der Analyse und Visualisierung im Besonderen besteht darin, die Optionen, Annahmen und möglichen Effekte von Entscheidungen zu exponieren. Visualisierung kann als Medium genutzt werden, um die Unsicherheit in diesem Prozess für einen kritischen Diskurs über

- die Beschränkungen der verwendeten Verfahren,
- die eigenen Annahmen,
- die Qualität der Resultate und
- die Reichweite der Implikationen

zu exponieren.
Ebenso wie die Wahl eines Data-Mining-Verfahrens determiniert die Wahl einer Visualisierung die Informationen, die daraus gewonnen werden können. Jede Visualisierung - gleich, ob sie ein Teilergebnis oder ein Endergebnis der Analyse darstellt - setzt verschiedene Größen miteinander in Bezug. Mindestens ebenso bedeutsam, wie das, was eine Visualisierung darstellt, ist jedoch das, was sie nicht darstellt; der implizite Ausschluss aller anderen Faktoren aus dem Fokus des Betrachters. Dies ist ebenso ein Problem der visuellen, explorativen Datenanalyse wie auch der Präsentation von Ergebnissen. In beiden Fällen schafft die Visualisierung einen Kontext, auf dem die folgenden Entscheidungen gründen - und engt diesen Kontext gleichzeitig ein.
Entscheidungsträger sollten ihre Entscheidungen auf der Basis der relevanten Argumente begründen und verteidigen, anstatt auf Grundlage der einfachen Argumente. Ausgangspunkt dieser Arbeit war daher die Entwicklung einer Visualisierungstechnik, die nicht nur viele Attribute eines Datensatzes gleichzeitig darstellen sollte, sondern die vor allem auch so viele Bezüge wie möglich zwischen diesen Attributen sichtbar machen sollte. Dabei zeigte sich, dass es zwar möglich ist, zehn- und mehrdimensionale Zusammenhänge als Muster sichtbar

[1] Hier, wie auch im Folgenden, soll bei personenbezogenen Substantiven die männliche Form verwendet werden. Die Kriterien für diese Entscheidung sind (in absteigender Priorität): Lesbarkeit, Begrifflichkeit (die männliche Form darf für Personengruppen beiderlei Geschlechts vereinnahmt werden), Gewohnheit und die Einsicht, dass man mit Grammatik Politik machen *kann*, aber keineswegs *muss*.

zu machen, ebenso deutlich wurden jedoch auch die Grenzen eines solchen Ansatzes: Auch wenn die Visualisierung darstellen konnte, *dass* Zusammenhänge existierten, bedeutet es stets einen hohen Lernaufwand, bis ein Anwender dazu in der Lage war, diese Zusammenhänge zu *lesen*.
Daraus ergab sich die Frage, ob eine Visualisierung nur höchstens Zusammenhänge darstellen „darf“, die gerade so komplex sind, dass ein Anwender sie noch mit vertretbarem Aufwand lesen kann. Gegen eine solche Einschränkung sprach und spricht, dass sich dann das genutzte Potential der menschlichen Mustererkennung auch an den kognitiven Fähigkeiten des Menschen orientieren muss. Dagegen spricht auch, dass Visualisierung prinzipiell die Möglichkeit bietet, komplexe Zusammenhänge so augenfällig zu machen, dass sie nicht ignoriert werden können - und zwar selbst dann, wenn sie nicht sofort erklärt werden können. Durch die in dieser Arbeit vorgestellten Methoden soll dem Menschen die Übersetzung bereitgestellt werden, um die Fähigkeiten der Wahrnehmung und die ebenso wichtigen kognitiven Fähigkeiten - innerhalb ihrer Domäne - optimal einzusetzen.

1.1 Ziele dieser Arbeit

In dieser Arbeit werden zwei neue Methoden für die Kopplung zwischen automatischen und visuell-interaktiven Verfahren für die Datenanalyse vorgestellt. Mit der ersten Methode soll die explorative Datenanalyse unterstützt werden: Die Suche nach Mustern und Strukturen in den Daten und deren formale Beschreibung in einem Modell oder einer Hypothese. Durch die zweite Methode soll die konfirmative Datenanalyse unterstützt werden: Die Bestätigung oder Widerlegung einer gegebenen Hypothese oder eines Modells durch einen Abgleich mit Referenzdaten. Die beiden Kopplungen zwischen automatischen und visuell-interaktiven Verfahren sind dabei Mittel für die Umsetzung von zwei der drei Hauptziele dieser Arbeit. Das dritte Ziel ist die Synthese zwischen den beiden Ansätzen.

Das erste Ziel für die Kopplung der Technologien ist eine Aufgabenteilung zwischen Mensch und Maschine, die den spezifischen Stärken und Schwächen beider gerecht wird. Sowohl Techniken aus den Bereichen der Informationsvisualisierung, wie auch aus dem Data-Mining, dienen der Suche und Beschreibung von nicht-trivialen Zusammenhängen in Daten.
In der Informationsvisualisierung übernimmt der Mensch diese Aufgaben. Eine Visualisierungstechnik soll dabei die Suche nach Zusammenhängen dadurch unterstützen, indem die Daten so dargestellt werden, dass die Fähigkeiten des Menschen zur Erkennung von Mustern, zur Segmentierung von Bildern und zur Separation von Rauschen optimal genutzt werden. Die Suche nach Mustern wird dadurch im Idealfall zu einem präattentiven Prozess, der nur wenig kognitive Ressourcen beansprucht. Für die Beschreibung eines Musters gilt dies jedoch nicht.
Im Kontext des wahrgenommenen Bilds ist das Muster jedoch für die Analyse bedeutungslos. Die Beschreibung eines Musters - die Umwandlung eines visuellen Artefakts in ein sprachliches Artefakt - ist notwendig, um die Informationen, die das Muster repräsentiert, im Kontext der Aufgaben und Ziele der Analyse interpretieren und bewerten zu können. Hierbei handelt es sich um einen kognitiven Prozess, der für jede neue Visualisierungstechnik, jede neue Konfiguration der Darstellung und schließlich jedes neue Muster neu durchgeführt werden muss. Die Herstellung der Korrespondenz zwischen wahrgenommenem Muster und der Analyseaufgabe beansprucht daher kognitive Ressourcen. Mit der hier vorgestellten Methode für die Kopplung soll der Mensch von dieser Aufgabe soweit wie möglich entlastet werden.
Im Data-Mining übernimmt ein automatisches Verfahren die beiden Aufgaben der Suche und Beschreibung von Mustern. Die Sprache, in der die Beschreibung konstruiert wird, ist dabei Teil des Verfahrens. Gemeinsam mit der Suchheuristik und Qualitätskriterien definiert diese Modellierungssprache, welche Teilmengen einer Grundmenge von Daten als Muster erkannt und beschrieben werden. Die Wahl eines Verfahrens und all seiner Parameter muss vor dem Analyseschritt getroffen werden, wobei das Ergebnis der Analyse schließlich nicht nur durch die Daten, sondern auch durch diese Auswahl beeinflusst wird.

Es gibt mehrere Fälle, in denen die automatische Suche von Mustern fehleranfällig ist:

- Die Daten enthalten einen hohen Anteil Rauschen, d.h. Informationen, die für die Konstruktion eines Modells irrelevant sind und ausgeschlossen werden sollen.
- Strukturen und Muster bilden keinen Kontrast zueinander oder zum Hintergrund. Ein Spezialfall ist der so genannte „Curse-of-Dimensionality“, der dann auftritt, wenn die Struktur der Datenverteilung durch die hohe Dimensionalität ihres Merkmalsraums dominiert wird.
- Verschiedene Strukturen des Datensatzes werden nur durch verschiedene Modelle „gut“ beschrieben. Da die Heuristiken von Data-Mining-Techniken Optimierungsverfahren sind, kann man erwarten, dass die Suche gegen ein lokales Optimum konvergiert. In diesem Fall wäre dieses Optimum jedoch lediglich ein Kompromiss, in dem ein Modell konstruiert wird, das die Strukturen nur im Mittel gut beschreibt.
- Verschiedene Strukturen des Datensatzes werden nur durch unterschiedliche Verfahren gut beschrieben. Dabei handelt es sich um die Verschärfung des eben genannten Falls.

In solchen Problemfällen ist der Mensch bei der Suche von Mustern automatischen Verfahren überlegen. Die Segmentierung eines Bilds gelingt auch unter Störeinflüssen, benötigt keine Voreinstellung und passt sich überdies den wahrgenommenen Mustern flexibel an. Daher für die Aufgabenteilung zwischen Mensch und Maschine eine Kopplung vorgeschlagen, in der der Mensch die Mustererkennung und Segmentierung des Bilds übernimmt, und durch ein automatisches Verfahren eine formale Beschreibung für ein von Menschen identifiziertes Muster zu konstruieren.
Aus der Sicht der Informationsvisualisierung bzw. der Nutzung ihrer Techniken verfolgt die Trennung von Mustererkennung und Musterbeschreibung ein abgeleitetes Ziel. Eine Annahme für das Konzept dieser Arbeit besteht darin, dass eine Visualisierungstechnik die beiden Aufgaben in unterschiedlichem Maße unterstützt. Im Realisierungskapitel dieser Arbeit wird beispielsweise eine Visualisierungstechnik vorgestellt, durch die mehrdimensionale Zusammenhänge zwar leicht zu identifizieren, aber nur mit hohem Aufwand zu beschreiben sind. Bei vergleichbaren Visualisierungstechniken ist es daher möglich, dass der potentielle Vorteil, der durch die Darstellung komplexer Zusammenhänge als Muster gewonnen wird, dadurch verloren geht, dass diese Zusammenhänge auch beschrieben werden müssen, bevor sie interpretiert werden können. Die Nutzung automatischer Methoden für die Beschreibung der wahrgenommenen Muster kann daher auch eine Strategie für die Verwendung von Visualisierungstechniken sein, die sonst nicht genutzt werden können.

Die Zielsetzung für die zweite Kopplung besteht darin, durch einen visuellen Abgleich eine Bewertung für die Qualität des Modells *und* eine Bewertung für die Qualität des Verfahrens, das dieses Modell erzeugt, zu ermöglichen. Damit soll das Problem entschärft werden, dass ein automatisches Verfahren sich nicht selbst bewerten kann. Ein automatisches Verfahren optimiert ein Modell nach vorgegebenen Gütekriterien. Diese Kriterien gehören zum Verfahren und müssen vorher festgelegt werden. Unter Umständen ist weder deren Bedeutung in Bezug auf die Aufgabenstellung der Analyse hinreichend gut bekannt, noch deren Effekt auf

die Konstruktion des Modells. Sie stehen daher selbst zur Disposition.
Die Gütekriterien für die Modellierung, wie auch statistische Kennzahlen für die Prädiktionsgüte eines Modells sind quantitative Maße, wie sie im allgemeinen auch als Kriterien für die Bestätigung oder Widerlegung einer Hypothese in der konfirmativen Analyse verwendet werden. Das Repertoire der Kriterien für die Bewertung eines Modells soll dabei durch einen visuellen Abgleich ergänzt werden. Für die Datenanalyse wird dabei ein Konzept übernommen, das aus der Wissenschaftlichen Visualisierung für physikalische Modelle seit langen bekannt ist: Die Konstruktion von Daten auf der Basis von prädiktiven Modellen durch Simulation und deren Abgleich mit den Referenzdaten in einer Visualisierung.
Der visuelle Abgleich soll qualitative Informationen über das Modell liefern. Insbesondere soll unterschieden werden können, ob der Prädiktionsfehler ein systematischer Fehler ist, der sich beim visuellen Abgleich als Muster manifestiert oder aber ob der Fehler lediglich ein Rauschen ist. Allein mit quantitativen Fehlermaßen ist eine solche Unterscheidung schwer oder sogar unmöglich.

Mit der ersten Kopplung wird eine Transformation von Daten über Muster hin zu Modellen für die explorative Analyse beschrieben. Mit der zweiten Kopplung wird eine Transformation in umgekehrte Richtung für die konfirmative Analyse beschrieben. Die Verbindung von explorativer und konfirmativer Analyse ist die dritte Zielsetzung dieser Arbeit. Sie beruht darauf, dass die vorgenannten Transformationen komplementär zueinander sind. Die Verbindung erzeugt einen zyklischen Prozess mit dem Ziel, eine belastbare Korrespondenz zwischen Daten, Mustern und Modellen herzustellen. Dieses Ziel gilt gleichermaßen für die konfirmative und die explorative Analyse.
Die explorative Analyse soll durch diese Verbindung insofern verbessert werden, da die Bewertung des Modells und des Verfahrens, das es konstruiert, durch den komplementären Prozess kontrolliert werden kann. Über diesen Prozess - die Simulation und den visuellen Abgleich - wird eine Feedbackschleife zwischen Mensch und Maschine geschlossen. Die Kontrolle über dieses Feedback wird möglich, da es unabhängig von den Parametern des Verfahrens funktioniert, das das Modell konstruiert. Diese Kontrolle soll es dem Anwender ermöglichen, zu entscheiden, ob ein Verfahren überhaupt geeignete Modelle für die Strukturen und Muster in den Daten konstruieren kann.
Die konfirmative Analyse soll durch diese Verbindung verbessert werden, in dem das Repertoire ihrer Verfahren erweitert wird, um jene Techniken, die für die explorative Analyse genutzt werden. Auf diese Weise wird das Repertoire quantitativer Kriterien für die Prädiktionsgüte ergänzt um qualitative Kriterien.

Diese Arbeit ist nicht die Erste, die sich mit der Kopplung von Techniken der Informationsvisualisierung und Data-Mining-Verfahren auseinandersetzt. Mit Abstand die meisten Systeme, in denen eine solche Verbindung etabliert wird, nutzen die Visualisierungstechniken als Medium für die Präsentation von (Zwischen-)Ergebnissen. Die Interpretation der dargestellten Informationen durch den Anwender im Kontext seiner Aufgabe ist der Ausgangspunkt für die weiteren Entscheidungen in der Analyse.
Im Unterschied zu diesen Ansätzen setzen die hier vorgestellten Methoden eine Interpretation des wahrgenommenen Bildes durch den Nutzer nicht voraus. Dies würde die im ersten Ziel dargestellte Arbeitsteilung zwischen Mensch und Maschine konterkarieren. Dabei muss jedoch betont werden, dass es nicht das Ziel ist, einen solchen Prozess für jede Aufgabe und

Entscheidung in der Analyse zu etablieren. Den Nachweis dafür zu erbringen, dass das möglich ist, würde den Rahmen einer Arbeit sprengen. Stattdessen setzen die Methoden voraus, dass die Aufgabe - die Herstellung der Beziehung zwischen Daten, Mustern und Modellen - klar umrissen, in den größeren Zusammenhang der Analyse eingebettet und dem Anwender bewusst ist. In diesem Sinne erweitern die hier vorgestellten Kopplungsvarianten das Repertoire von Methoden der konfirmativen und explorativen Datenanalyse für diese Aufgabe.
Abgesehen von den Verfahren, in denen Visualisierungstechniken für die Darstellung von Ergebnissen genutzt werden, werden in der Literatur eine Reihe weiterer Kopplungen vorgestellt, die sich nach Anwendungsgebiet, den eingesetzten Techniken oder aber nach den Ansatzpunkten der Kopplung unterscheiden. Ein weiteres Ziel dieser Arbeit ist daher die Abgrenzung verschiedener Varianten unterschieden nach den Ansatzpunkten der Kopplung innerhalb der Informationsvisualisierung und im Data-Mining - unter anderem vor dem Hintergrund, die hier vorgestellten Methoden einordnen zu können.
Aus diesem Grund sind jene Modelle der beiden Technologien, in denen die Verfahren allgemein beschrieben werden, von besonderem Interesse. Die Möglichkeiten für eine Verbindung von Informationsvisualisierung und Data-Mining sollen systematisch erfasst werden im Bezug auf die möglichen Ansatzpunkte der Verbindungen, die Menge der möglichen Kombinationen und die Richtung von Daten- und/oder Kontrollfluss.
Die Systematisierung der möglichen Kopplungsvarianten hat einen weiteren Grund. Die gemeinsame Nutzung von Visualisierungstechniken und automatischen Verfahren ist eine „Kopplung“ zwischen Techniken, der Art mit der Menschen seit jeher mit unterschiedlichen Werkzeugen auf den Werkstücken (oder eben Daten) operieren. Solange die Werkzeuge dabei unabhängig voneinander eingesetzt werden, handelt es sich nicht um eine technische, sondern um eine methodische Kopplung. Die Systematisierung der technischen Kopplungsvarianten dient daher auch als Ausgangspunkt dafür, die technischen Möglichkeiten der visuellen Datenanalyse von den rein methodischen Möglichkeiten voneinander abzugrenzen.

Zuletzt sollen für das Konzept noch zwei Anforderungen formuliert werden. Wie bereits beschrieben sollen die Kopplungsvarianten weder nach dem Anwendungsgebiet, noch nach den jeweils miteinander verbundenen Techniken kategorisiert werden. Auch wenn im Kapitel 4 eine konkrete Umsetzung des Konzepts beschrieben wird, sollen die Methoden im Konzept unabhängig von verschiedenen Techniken beschrieben werden. Ein Grund dafür ist die Verallgemeinerbarkeit des Konzepts für verschiedene Visualisierungstechniken und Verfahren des Data-Mining, die die Voraussetzung dafür ist, dass verschiedene Verfahren und Methoden bei der Analyse miteinander vergleichbar und gegeneinander austauschbar sind.

1.2 Gliederung der Arbeit

Die Arbeit ist wie folgt gegliedert: Das folgende Kapitel 2 *Wissenschaftliche Einordnung* ist in sechs Abschnitte unterteilt. Der erste Abschnitt 2.1 *Datenanalyse* behandelt allgemeine Zielsetzungen und Methoden der Datenanalyse. Die Datenanalyse ist selbst keine einheitliche wissenschaftliche Disziplin, sondern integriert in der Praxis Ansätze und Theorien aus sehr unterschiedlichen Disziplinen.
Diese Arbeit behandelt nicht alle Methoden und Theorien der Datenanalyse, sondern beschränkt sich auf Verfahren aus zwei wissenschaftlichen Disziplinen - aus der Informationsvisualisierung und aus dem Data-Mining. Auch behandelt diese Arbeit in erster Linie zwei, wenngleich sehr allgemein formulierte, Aufgaben: Eine dieser Aufgaben besteht darin, eine Korrespondenz herzustellen, zwischen den der Analyse zugrundeliegenden Daten, den Strukturen und Mustern, sowie formalen Modellen. Diese Aufgabe entspricht den im vorigen Abschnitt genannten ersten beiden Hauptzielen dieser Arbeit. Die zweite Aufgabe besteht darin, Methoden bereitzustellen, mit denen die Qualität dieser Korrespondenz unabhängig von den Verfahren, mit denen sie konstruiert wird, bewertet werden kann. Diese Aufgabe entspricht dem dritten Hauptziel dieser Arbeit. Ein Ziel dieses Abschnitts besteht darin, diese Aufgaben innerhalb der Analyse zu lokalisieren und auch von anderen Forschungsdisziplinen abzugrenzen.
Das zweite Ziel dieses Abschnitts ist es, die allgemeinen Ziel- und Problemstellungen in der Datenanalyse zu beschreiben, die unabhängig von bestimmten Disziplinen und Verfahren sind. Zu den betrachteten Problemstellungen zählt insbesondere, auf welche Weise sichergestellt werden kann, dass die Analyseverfahren nach einem methodisch sicheren Ansatz gewählt und genutzt werden. Allgemein bedeutet dies, dass explizit beschrieben werden muss, inwiefern das Ergebnis einer Analyse abhängt von den zugrundeliegenden Daten und inwiefern das Ergebnis abhängig ist von der Wahl eines bestimmten Analyseverfahrens. Dieses allgemeine Problem wird in den folgenden beiden Abschnitten dieses Kapitels übertragen auf die Auswahl und Steuerung automatischer Verfahren der Analyse.
Das dritte Ziel dieses ersten Abschnitts besteht darin, die Gemeinsamkeiten und Unterschiede zwischen explorativer und konfirmativer Datenanalyse zu beschreiben. Für die Zusammenführung der beiden Strategien im Konzept dieser Arbeit soll so eine gemeinsame Grundlage geschaffen werden.

Die folgenden beiden Abschnitte behandeln automatische Verfahren der Datenanalyse, wobei Data-Mining und Machine-Learning-Verfahren zusammengefasst werden. Abschnitt 2.2 *Automatische Verfahren der Datenanalyse - Knowledge Discovery in Databases (KDD)* stellt das Modell für den KDD-Prozess (Fayyad et al. [FPSS96b]) vor. Dieses Modell dient der Einordnung des nächsten Abschnitts, welcher sich mit dem Data-Mining und Machine-Learning selbst befasst, in den größeren Rahmen des Knowledge-Discovery-Prozess. Der KDD-Prozess umfasst alle Teilschritte der Datenanalyse und beschreibt insbesondere auch die Ansatzpunkte für Entscheidungen über Verfahren und Methoden durch den Menschen. Von allen möglichen Freiheitsgraden der Analyse, über die der Anwender entscheiden kann, werden zwei im Konzept dieser Arbeit behandelt: Die Bewertung und Auswahl eines Data-Mining Verfahrens und dessen Steuerung. Inwiefern das Konzept dieser Arbeit auch auf andere Freiheitsgrade in der Analyse übertragen werden kann wird im Ausblick der Arbeit diskutiert.

Der folgende Abschnitt 2.3 *Automatische Verfahren - Data-Mining und Machine-Learning* widmet sich diesem Teilschritt des KDD-Prozesses. Dieser enthält zwei der zentralen Teile dieses Kapitels: In Abschnitt 2.3.3 *Allgemeine Beschreibung von Data-Mining-Verfahren* werden Data-Mining Verfahren in Bezug auf die Funktion, die sie im Analyseprozess erfüllen, unabhängig von einer bestimmten Aufgabe beschrieben. Auch die in der Literatur identifizierten Grenzen bei der Anwendung automatischer Verfahren werden in diesem Abschnitt genannt. Diese motivieren die Kopplung zwischen automatischen und visuell-interaktiven Methoden aus der Sicht des Data-Mining.
In Abschnitt 2.3.4 *Komponenten von Data-Mining-Verfahren*, wird ein Modell vorgestellt, das die Komponenten beschreibt, die allen Data-Mining (und Machine-Learning) Verfahren im Wesentlichen gemeinsam sind. Anhand dieses Modells werden die drei verschiedenen Ansatzpunkte für die Steuerung automatischer Verfahren identifiziert, die die Grundlage sind für die Systematisierung verschiedener Kopplungsvarianten am Ende des Kapitels.
Die Beschreibung der automatischen Methoden beendet ein Überblick über verschiedene Verfahren in Unterabschnitt 2.3.5. Diese werden abhängig von den unterstützten Aufgaben der Datenanalyse eingeordnet. Dabei werden Entscheidungsbäume im Detail behandelt, um die verschiedenen Ansatzpunkte für die Kopplung beispielhaft vorzustellen und weil diese Verfahren in der Realisierung des Konzepts verwendet werden.

Abschnitt 2.4 *Informationsvisualisierung* behandelt die Modelle der Informationsvisualisierung. Darin wird zunächst die allgemeine Definition vorgestellt, und anschließend wird der Fokus dieses Abschnitts auf die Informationsvisualisierung für die explorative Datenanalyse eingegrenzt. Dies steht im Gegensatz zu Verfahren, die in erster Linie für die Präsentation verwendet werden. Diese Unterscheidung ist vor dem Hintergrund relevant, da das erste Ziel dieser Arbeit - die Trennung von Mustererkennung und Musterbeschreibung - vor allem relevant ist für die Techniken, die für die visuelle explorative Datenanalyse verwendbar sein sollen.
In Abschnitt 2.4.2 wird der Informationsvisualisierungs-Prozess (Card et al. [CMS99]) vorgestellt, mit dem analog zum Modell für die automatischen Verfahren, die Ansatzpunkte für eine Kopplung identifiziert werden können. Der technischen Beschreibung des Visualisierungsprozesses wird die Taxonomie der Aufgaben (Amar und Stasko [AS05]) gegenübergestellt. Diese zweite Perspektive, in der die Nutzung einer Visualisierung an den Aufgaben orientiert wird anstatt an den vorliegenden Daten, motiviert die Möglichkeit eine Visualisierungstechnik auf die Aufgabe der Mustererkennung zu fokussieren.
Die folgenden beiden Unterabschnitte 2.4.3 *Interaktion* und 2.4.4 *Visuelle Wahrnehmung* stellen weitere Grundlagen für das Konzept dieser Arbeit vor. In beiden Unterabschnitten werden Modelle vorgestellt, anhand derer die Mustererkennung und die Musterbeschreibung als getrennte Prozesse beschrieben werden; dies wird beispielsweise im Informationsvisualisierungsprozess nicht explizit deutlich. Diese Trennung wird im Konzept einerseits auf der Grundlage des Aktionsstufenmodells (Norman [Nor02] bzw. Lam [Lam08]) beschrieben und andererseits auf der Grundlage des dreistufigen Wahrnehmungsmodells von Ware [War04b].
Zwei Verfahren für die Interaktion werden im Detail vorgestellt: Die *Direkte Selektion* und *Linking & Brushing*. Die Konzepte für beide Interaktionsverfahren können unabhängig von einer bestimmten Visualisierung beschrieben werden. Die technischen Aspekte der im Konzept vorgestellten Kopplungsvarianten werden als Verallgemeinerung dieser Verfahren dargestellt.

Der vorletzte Abschnitt des Kapitels beschreibt schließlich das Forschungsgebiet 2.5 *Visual Analytics*, mit dem Fokus auf das *Visual Data Mining*. Es behandelt dabei insbesondere die existierenden Kopplungsmodelle für die Verbindung von automatischen und visuell-interaktiven Verfahren der Datenanalyse. Das Modell für den *Visual Analytics Prozess* (Keim et al. [KAF+08]) ist die Grundlage für die Einordnung der in den vorangegangen Abschnitten identifizieren Ansatzpunkte. Die verschiedenen Kombinationen werden dabei, wenn möglich, auch bereits existierenden Ansätzen zugeordnet. Zwei dieser Kombinationen machen den Schwerpunkt des Konzepts dieser Arbeit aus. Sie beschreiben die Umsetzung der ersten beiden Hauptziele dieser Arbeit. Zwei weitere Kopplungsvarianten werden ebenfalls in das Konzept integriert: Die Erste - die Visualisierung von Modellen - ist nicht neu; sie vervollständigt jedoch das Konzept, da es die Möglichkeit des visuellen Abgleichs auch auf der Ebene der Modelle bietet. Die zweite dieser Varianten - die Parametrisierung automatischer Verfahren aus einer Visualisierung - wird beschrieben, weil sie bisher nur in sehr wenigen Ansätzen Verwendung findet. Eine ausführliche Diskussion der Perspektiven, die eine solche Kopplung bietet, liegt jenseits des Rahmens dieser Arbeit. Mögliche Anknüpfungspunkte werden daher im Ausblick aufgegriffen.

Das folgende Kapitel 3 *Konzept* ist unterteilt in drei Abschnitte. Der erste Abschnitt ist dem ersten Ziel dieser Arbeit gewidmet und beschreibt die Kopplungsvarianten, die sich durch die Trennung von Mustererkennung und Musterbeschreibung ergeben. Dabei sind jeweils die Visualisierungstechniken die Quelle und die Data-Mining-Verfahren das Ziel des Datenflusses in der Verbindung. In diesem Abschnitt wird der Weg des Daten- bzw. des Kontrollflusses von der Wahrnehmung in der Visualisierung über die Interaktion bis hin zur Steuerung des automatischen Verfahrens nachgezeichnet. Die Unterabschnitte 3.1.1 *Rolle von Mustern im Data-Mining und in der Informationsvisualisierung* und 3.1.3 *Interaktion als Datenquelle* motivieren dabei die Trennung zwischen Mustererkennung und Musterbeschreibung aus der Sicht des Nutzers.
Unterabschnitt 3.1.4 *Nutzung der Daten aus der Interaktion in automatischen Verfahren* beschreibt die Schnittstelle zwischen Visualisierungstechniken und den automatischen Verfahren. Darin wird beschrieben, auf welche Weise die durch die Interaktion definierbaren Informationen für die Steuerung der automatischen Verfahren verwendet werden können. Grundlage für diesen Abschnitt ist die vorgestellte Systematik.
Die folgenden Abschnitte behandeln verschiedene Aufgabenstellungen im Data-Mining und sollen zeigen, wie die verschiedenen Kopplungsvarianten für diese Aufgaben genutzt werden können, und auch wie sie bereits genutzt werden. Dabei wird auch die Verbindung, die im Kapitel 4 exemplarisch umgesetzt wird, allgemein beschrieben, und im Rahmen der Systematik mit anderen bereits existierenden Verfahren verglichen.
Eine Sonderstellung nimmt der Unterabschnitt 3.1.5 *Modellvisualisierung* ein. Es handelt sich dabei nicht um eine Kopplung von einer Visualisierung zu einem automatischen Verfahren; vielmehr ist der Datenfluss umgekehrt. Ebenso ist diese Art der Kopplung bereits vielfach umgesetzt. Sie ist aus dem Grund Teil des Konzepts, weil sie den Endpunkt des Prozesses darstellt, in dem die Inhalte dem Anwender für die Interpretation und Bewertung auf einem höheren Abstraktionsniveau - auf der Ebene des Modells - zugänglich gemacht werden.
Im Unterabschnitt 3.1.6 *Zusammenfassung der Systematik* werden die vorgestellten Kopplungsvarianten verglichen. Zudem wird die Motivation für die Trennung von Mustererken-

nung und Musterbeschreibung aus der Sicht der Informationsvisualisierung (bzw. des Anwenders) ergänzt um eine entsprechende Motivation aus der Sicht eines Data-Mining Verfahren. Dieser Abschnitt nimmt Bezug auf die in Kapitel 2.3.3 vorgestellten Defizite der automatischen Verfahren.

Während der erste Teil des Konzepts die Konstruktion von Modellen aus Daten behandelt, behandelt der zweite Teil 3.2 *Konfirmatives Feedback* den umgekehrten Weg und damit das zweite und dritte Ziel der dieser Arbeit. Für die Überprüfung von prädiktiven Modellen werden diese durch Simulation in Daten transformiert. Dieser Prozess kann in zwei Situationen genutzt werden, abhängig davon, woher das Modell stammt.
In der ersten Variante, beschrieben in Unterabschnitt 3.2.1 *Konfirmative visuelle Analyse für prädiktive Modelle*, ist das Modell der Ausgangspunkt einer konfirmativen Analyse. Durch die Transformation von Modellen in Daten soll der visuelle Abgleich zwischen simulierten Daten und Referenzdaten ermöglicht werden, um das Repertoire für die konfirmative Analyse zu ergänzen.
In der zweiten Variante, beschrieben in Unterabschnitt 3.2.3 *Visuelles Feedback für die explorative Analyse* stammt das Modell von einer vorangegangenen Datenanalyse, wie sie im ersten Teil des Konzepts beschrieben wird. In diesem Abschnitt wird ebenfalls beschrieben, wie die beiden komplementären, jedoch unabhängigen, Prozesse miteinander gekoppelt werden, um das dritte Hauptziel dieser Arbeit zu erreichen. Durch den visuellen Abgleich der Muster in den Referenzdaten und der Muster in den simulierten Daten wird untersucht, wie die beiden automatischen Verfahren visuell gegeneinander validiert werden können.
Der letzte Abschnitt des Konzepts 3.3 *Zusammenfassung* stellt die Varianten der Kopplung noch einmal zusammen. Zudem werden die verschiedenen Kopplungsvarianten in dem zyklischen Prozess zusammengefasst, der für die konfirmative und die explorative Analyse genutzt werden soll.

Die exemplarische Umsetzung des Konzepts wird im Kapitel 4 *Realisierung* beschrieben. Als Ausgangspunkt der Realisierung wird eine Visualisierungstechnik für die Darstellung mehrdimensionaler Daten und Bezüge vorgestellt. An diesem Beispiel wird auch die Diskrepanz zwischen Mustererkennung und Musterbeschreibung - und deren Unterstützung in der Informationsvisualisierung - noch einmal illustriert.
Die Visualisierungstechnik basiert auf den *Karnaugh-Veitch-Diagrammen* [Kar53], die ursprünglich für die Optimierung Boolescher Schaltungen entwickelt wurden. Diese Visualisierung ermöglicht es, Muster darzustellen, die mehrdimensionale Bezüge zwischen den Attributen eines Datensatzes darstellen. Die automatische Beschreibung der wahrgenommenen Muster wird in diesem Beispiel durch Entscheidungsbäume modelliert.
In der Darstellung der Kopplungsvarianten folgt dieses Kapitel der gleichen Reihenfolge wie das Konzept: Zuerst wird dargestellt, wie die Trennung von Mustererkennung und Musterbeschreibung umgesetzt werden kann. Anschließend wird das iterative Feedback, sowie der visuelle Abgleich für die Verifikation von Hypothesen bzw. die Validierung der verwendeten automatischen Verfahren vorgestellt.
Im letzten Kapitel 5 *Zusammenfassung und Ausblick* werden die in diesem Kapitel benannten Ziele aufgegriffen und diskutiert. Darüber hinaus werden zwei Forschungsfragen identifiziert, die auf dem Konzept dieser Arbeit aufbauen, und über die das hier vorgestellte Konzept erweitert bzw. verallgemeinert werden kann.

Kapitel 2

Wissenschaftliche Einordnung

2.1 Datenanalyse

In der *Forschungsagenda für Visual Analytics* beschreiben Thomas & Cook [TC05] folgende Zielsetzung für die Datenanalyse: Grundsätzlich soll das Ergebnis einer Analyse einem Menschen ermöglichen, eine Situation besser zu beurteilen und damit belastbare Entscheidungen treffen zu können. Ein Urteil beruht teilweise auf Informationen, die die spezifische Situation beschreiben und dem Menschen (vor dem Eintreten der Situation) nicht bekannt sind. Zusätzlich gründet ein Urteil auch auf dem Vorwissen des Menschen über die Anwendungsdomäne und darüber hinaus, auf Annahmen, auf dem Situationsverständnis und vor allem den Zielen, die mit der Entscheidung verfolgt werden sollen. Für eine Entscheidung müssen diese Inhalte miteinander verknüpft werden. Die Analyse nimmt die Entscheidung nicht ab. Das Ergebnis einer Analyse soll stattdessen helfen,

- die Beziehungen zwischen den die Situation beschreibenden Inhalten zu kennen und erklären zu können,
- Entscheidungsoptionen aus diesen Beziehungen und den Entscheidungszielen zu entwickeln,
- die Konsequenzen möglicher Entscheidungen einschätzen zu können.

Der Ausgangspunkt der Analyse ist dabei immer eine Fragestellung, die aus einem konkreten Anwendungsszenario abgeleitet, aber auch aus reinem Interesse verfolgt werden kann. Die Datenanalyse ist ein Prozess, in dem Inhalte zwischen verschiedenen Abstraktionsstufen übersetzt werden.

Thomas & Cook nennen *(ebd.)* diese Inhalte *analytische Artefakte* und unterscheiden dabei vier verschiedene Stufen:

- *Elementare Artefakte:* Artefakte, die einzelne Informationen darstellen, die bezüglich ihrer potentiellen Relevanz zur Fragestellung ausgewählt wurden. Thomas & Cook beschreiben elementare Artefakte als „isolierte“ Informationen. Dies soll hier jedoch so abgeschwächt werden, dass durchaus Beziehungen zwischen den Informationen bekannt sein können. Tatsächlich muss dies sogar der Fall sein, da es sonst nicht möglich wäre, die Informationen miteinander zu verknüpfen. Allerdings seien die bekannten Beziehungen entweder trivial und/oder irrelevant für die Fragestellung. Die Identifikation neuer Beziehungen ist eines der Ziele der Analyse. Elementare Artefakte aus externen Quellen können Messwerte, erhobene Daten, Dokumente etc. sein. Andere Quellen sind Annahmen des Analysten, oder auch die Analyse selbst - etwa in der Form von Simulationsdaten.

- *Musterartefakte:* Artefakte, die „nicht-beliebige“ Mengen elementarer Artefakte zusammenfassen. Was „nicht-beliebig“ bedeutet, wird durch die in der Analyse eingesetzten Modelle und Verfahren implizit oder explizit definiert. Allgemein soll „nicht-beliebig“ zunächst bedeuten, dass die Menge von potentieller Relevanz für die Beantwortung der Fragestellung ist. Muster erreichen ein höheres Abstraktionsniveau als elementare Artefakte, wenn deren Ähnlichkeiten oder auch Unterschiede effizienter beschrieben werden können, als es durch die Auflistung ihrer Elemente möglich wäre. Muster sind die Grundlage allgemeiner Aussagen, die mit den folgenden Artefakten konstruiert werden.

- *Artefakte höherer Ordnung:* Thomas & Cook zählen zu diesen Artefakten unter anderem logische Schlussfolgerungen und Modelle. Mit diesen Artefakten werden Verknüpfungen zwischen anderen Artefakten hergestellt, insbesondere auch solcher, die auf verschiedenen Abstraktionsebenen liegen. Ein Modell ist die formale Repräsentierung von Beziehungen zwischen Artefakten, die innerhalb der Analyse konstruiert und/oder bestätigt werden. Die Anwendung von Modellen überspannt alle Abstraktionsebenen der Analyse: Sie werden entwickelt für die Beschreibung von Daten, die Beschreibung von Mustern bzw. dafür wie Daten zu Mustern zusammengefasst werden können, aber auch für die Beschreibung der Fragestellung und Voraussagen der Folgen möglicher Entscheidungen.

- *Artefakte höchster Ordnung (orig.: “Complex reasoning constructs“):* Bei diesen Artefakten handelt es sich um jene, deren Abstraktionsstufe der der Analyse zugrundeliegenden Fragestellung entspricht. Dazu gehören aus der Fragestellung abgeleitete Hypothesen, die in der Analyse untersucht werden, ebenso wie die Präsentation der Ergebnisse in einer Form, die an den Anwendungskontext angepaßt ist. Das Ergebnis der Analyse ist einerseits die Beantwortung der Fragestellung, ggf. aber auch die Rückführung der Argumentation auf die zugrundeliegenden Fakten oder die Begründung der Analysemethodik.

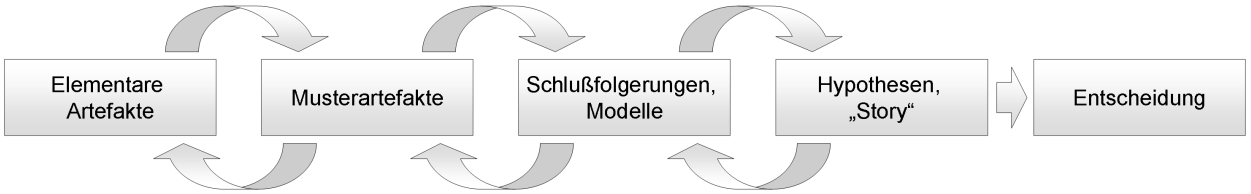


Abbildung 2.1: *Die Datenanalyse ist ein Prozess, in dem eine Verbindung zwischen Daten und Entscheidungen konstruiert wird. Jedes Analyseverfahren beschreibt dabei eine Transformation eines analytischen Artefakts in ein anderes. Bei der Konstruktion dieser Verbindung muss darauf geachtet werden, dass der Gehalt der Informationen zwischen den einzelnen Abstraktionsstufen nicht verfälscht oder verzerrt wird.*

Thomas & Cook beschreiben den Prozess des analytischen Schließens als die Konstruktion einer Verbindung zwischen jenen Artefakten, deren postulierte Korrektheit die Grundlage der Schlussfolgerungen darstellt und jenen Artefakten, deren Korrektheit in der Analyse zur kritischen Disposition steht (siehe Abbildung 2.1). Das bedeutet auch, dass für die Analyse einer gegebenen Fragestellung für bestimmte Artefakte die Korrektheit postuliert werden *muss*. Zu diesen Artefakten gehören typischerweise externe Datenquellen, aber auch das Wissen des Analysten über den Einsatz und die Steuerung von Analyseverfahren. Ohne eine Referenz von als „wahr“ erachteten Artefakten sind analytische Schlussfolgerungen nicht möglich.
Die Tatsache, dass prinzipiell alle Artefakte - Hypothesen, aber auch Daten und Annahmen - zur kritischen Disposition stehen können, steht dazu nicht im Widerspruch. Entscheidend ist, dass innerhalb einer Analyse nicht grundsätzlich *alle* Informationen hinterfragt werden können. Um überhaupt zu Ergebnissen zu gelangen muss mit der Fragestellung auch ein Referenzrahmen festgelegt werden, gegen den die Ergebnisse geprüft werden können. Die Ergebnisse sind dann in diesem Referenzrahmen gültig. Eine Analyse kann wertlos werden, sobald nur eine ihrer Referenzen kompromittiert wird. In vielen Fällen erfordert die Analyse einer Fragestellung daher die stufenweise Eliminierung von Unsicherheiten, die jeweils in der Formulierung neuer Fragestellungen münden. Die kann im Einzelfall bedeuten, dass für deren Analyse wiederum neue Artefakte aus unabhängigen Quellen als Referenz gesucht werden müssen.
Eine zentrale Problematik in der Datenanalyse besteht nicht allein darin, Unsicherheiten in der Schlusskette zu eliminieren, sondern sie überhaupt erst zu finden. Thomas & Cook fordern *(ebd.)* daher Transparenz und Belastbarkeit als zentrale Eigenschaften des Analyseprozesses. Bei der Analyse lässt sich eine Vielzahl einzelner Entscheidungen des Analysten identifizieren: Die Auswahl von Referenzdaten, die Konstruktion und Wahl von Modellen, Argumenten oder Verfahren und deren Steuerung. Es herrscht ein breiter Konsens darüber, dass Datenanalyse ein nichtlinearer Prozess ist [TC05, WB98, FPSS96b]. Insbesondere ist Datenanalyse kein rein induktiver Prozess, in dem ausgehend von elementaren Artefakten, d.h. von beobachtbaren und als korrekt erachteten Daten stets stufenweise abstraktere Artefakte konstruiert werden würden.
Die Analyse läuft deshalb nicht linear ab, weil man zum einen nicht davon ausgehen kann, dass ein Analyst von vornherein immer die „richtigen“ Entscheidungen fällt. Dies würde ein Verständnis für die Fragestellung voraussetzen, das durch die Analyse erst gewonnnen werden soll. Zum anderen ist es während einer Analyse häufig nicht möglich, eine Entscheidung

direkt nach dem Analyseschritt zu bewerten, in dem sie getroffen wird. Daraus folgt, dass bei der Analyse auch Entscheidungen getroffen werden müssen, bevor sie ausreichend begründet werden können. Belastbarkeit bedeutet dann, dass jede Entscheidung so behandelt werden könnte, als wäre sie *potentiell* falsch. Dass ein Analyseprozess im Allgemeinen iterativ abläuft ist letztlich ein Symptom dafür, dass Entscheidungen für die Analyse verglichen, verfeinert und gegebenenfalls revidiert werden müssen. Prinzipiell könnte jeder Schritt in

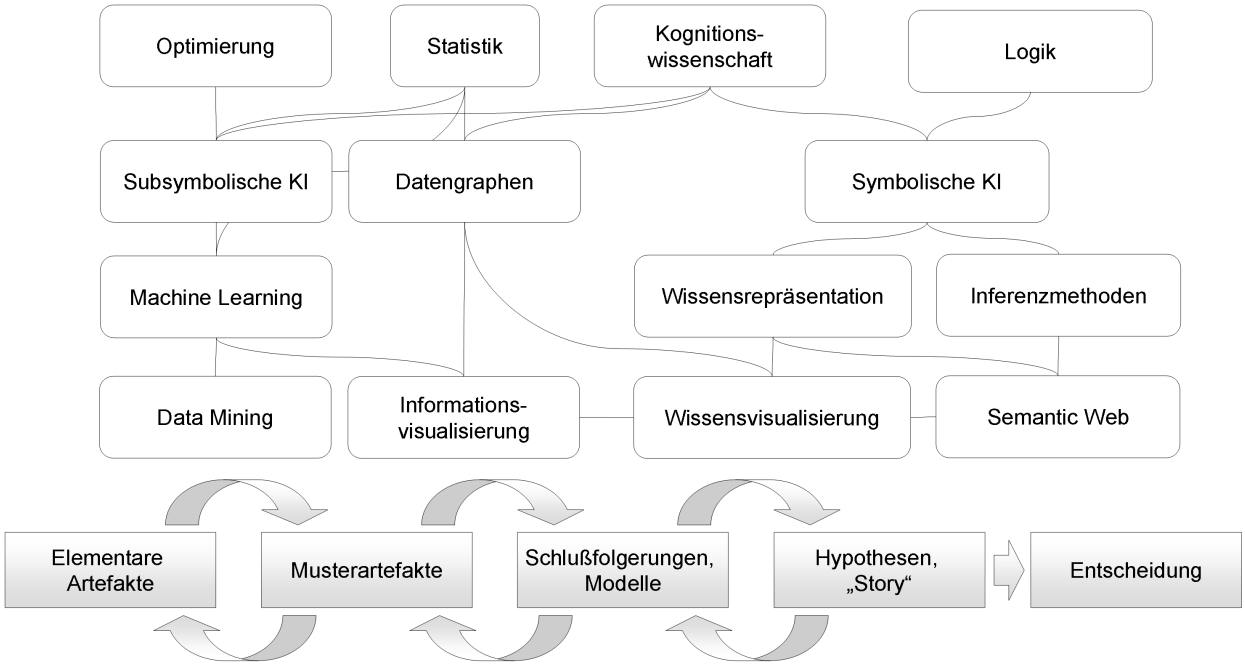


Abbildung 2.2: *Der Prozess des analytischen Schließens wird durch eine Vielzahl unterschiedlicher Forschungsgebiete behandelt. Hier sind nur die wichtigsten Verwandtschaftsbeziehungen zwischen diesen Gebieten dargestellt. In dieser Arbeit wird unterschieden zwischen wissensbasierten Technologien und „datenbasierten“ Technologien. Wissensbasierte Technologien gründen auf einer semantischen Beschreibung der Fakten in Form geeigneter Begriffe und Beziehungen. Datenbasierte Technologien gründen auf einer technischen Beschreibung der durch die Daten repräsentierten Fakten. Eine Schnittstelle zwischen den Technologien ergibt sich daraus, dass die Suche nach Strukturen und Beziehungen in datenbasierten Systemen die empirische Grundlage für die Begriffsbildung und die Weiterverarbeitung in wissensbasierten Systemen darstellt.*

der Kette auch alleine durch den Menschen durchgeführt werden. Menge und Komplexität der zu verarbeitenden Informationen machen jedoch meist eine maschinelle Verarbeitung notwendig. Für die Repräsentierung der Artefakte und deren Transformation zwischen den einzelnen Abstraktionsebenen werden Technologien aus sehr unterschiedlichen Forschungsgebieten eingesetzt (siehe Abbildung 2.2). Im Folgenden wird ein kurzer Überblick gegeben, in dem diese Forschungsgebiete innerhalb des Repertoires für die Datenanalyse lokalisiert werden können. Dies soll helfen, die technologischen Grundlagen und den Schwerpunkt dieser Arbeit stärker einzugrenzen.

- Fayyad et al. [FPSS96a] definieren das Ziel des *Knowledge-Discovery* und des *Data-Mining* als die Identifizierung nicht-trivialer, relevanter Muster aus Datenbeständen

oder die Beschreibung von Daten durch analytische Modelle. Die technische Domäne der Methoden sind daher die Transformation elementarer Artefakte in Muster bzw. in Modelle.

- Das Forschungsgebiet des *Machine-Learning* lässt sich anhand der Techniken vom Data-Mining nicht scharf abgrenzen. Machine-Learning Methoden dienen ebenfalls der Identifikation von Mustern in den Daten, wenngleich der Schwerpunkt nicht auf der Schöpfung von Wissen, sonder eher auf der Optimierung von automatischen Prozessen (z.B. für Textklassifikation, Bilderkennung, Robotersteuerung) liegt.

- Ziel der *deskriptiven Statistik* (siehe z.B. bei Lehn und Wegmann [LW06]) ist die Suche nach geeigneten Beschreibungen großer Datenbestände, die den Informationsgehalt der zugrundeliegenden Daten erhalten. Statistische Kenndaten werden ebenso auch in Data-Mining und Machine-Learning Verfahren verwendet. Methoden der *schließenden Statistik* gehören zum wichtigsten Repertoire für die Hypothesentests in der konfirmativen Datenanalyse.

- *Simulation* und *Vorhersage* sind Analysemethoden, in denen keine Abstraktion von zugrundeliegenden Daten hin zu Modellen und Hypothesen erfolgt; die Transformation der Artefakte erfolgt stattdessen in umgekehrte Richtung: Auf der Basis von Modellen werden elementare Artefakte konstruiert. Dies gilt allgemein sowohl für die physikalisch basierte Simulation, als auch für die Vorhersage abstrakter Daten. Es handelt sich dabei um indirekte Methoden für den Test von Modellen.

- Der Forschungszweig der *Wissensbasierten Systeme* umfasst grob eine Reihe von unterschiedlichen Disziplinen, die sich seit den achtziger Jahren aus den Zielstellungen der so genannten „schwachen künstlichen Intelligenz“ entwickelten (siehe u.a. Shadbolt und Milton [SM99] oder Crubézy und Musen [CM04]). Nach Svatek et al. [SLV04] lassen sich etwa für die Disziplin des *Knowledge Engineering* mehrere Phasen unterscheiden. Bis in die frühen neunziger Jahre dominierte der Einfluß der klassischen Forschung zur Künstlichen Intelligenz (KI), deren Ziele darin bestanden, Expertenwissen „abzuschöpfen“ (*Knowledge Akquision*) und in eine maschinell verarbeitbare Form zu bringen (siehe auch Wielinga et al. [WSB92]). In den neunziger Jahren begann die Verbreitung von Domain-Ontologien als „lingua franca“ der wissensbasierten Systeme. Die Domäne dieser Disziplinen ist der maschinelle Repräsentierung und Verarbeitung von Artefakten höherer und höchster Ordnung.

- *Informationsvisualisierung* definieren Card el al. [CMS99] als die Nutzung von computerunterstützten, interaktiven, visuellen Repräsentierungen abstrakter Daten für die Identifikation und Mustern und Zusammenhängen. Dieses Forschungsgebiet umfasst die visuell-interaktiven Verfahren für die Datenanalyse. Deren Modelle beschreiben die zweite Schnittstelle für die Kopplung, die im Rahmen dieses Konzepts entworfen wird. Informationsvisualisierung wird daher im Abschnitt 2.4 eingehend behandelt.

- *Wissensvisualisierung* ist verwandt mit der Informationsvisualisierung, bezieht sich aber auf die Darstellung von Modellen oder Hypothesen. Nach Burkhard et al. [Bur04] sind die Unterschiede zwischen Informations- und Wissensvisualisierung nicht technisch begründet, sondern in erster Linie durch ihre Zielsetzung.

Im Prinzip kann der Prozess des analytischen Schließens über alle Abstraktionsebenen hinweg technisch unterstützt werden. Die meisten Technologien beziehen sich dabei jedoch nur auf einen bestimmten Teil des Prozesses.

2.1.1 Datenanalyse und die „Analyse der Analyse“

Die Forderung nach belastbaren Entscheidungen muss bei der Analyse auf zwei Ebenen erfüllt werden. Die erste Ebene sind dabei die Artefakte selbst. Potentiell besteht immer die Gefahr, dass sich Unsicherheiten innerhalb des gesamten Analyseprozesses fortpflanzen. Eine Analyse auf der Basis potentiell unsicherer Daten hat im Prinzip keine Aussagekraft, wenn die Reichweite der Unsicherheit über den gesamten Prozess bis hin zu Ergebnis nicht ausreichend abgeschätzt werden kann und damit nicht kontrollierbar bleibt. Im Begriff „Datenqualität“ fassen Thomson et al. [THM+05] mehrere Merkmale wie unter anderem Genauigkeit, Fehler, Vollständigkeit, Konsistenz, Herkunft oder Glaubwürdigkeit der Quelle zusammen. Das Ziel der Analyse besteht darin, die Auswirkungen der Unsicherheit auf das Ergebnis zu minimieren *(ebd.)*.
Um die Propagation von Unsicherheiten in der Analyse abschätzen zu können, müssen die Verfahren zusammen mit den eigentlichen Anwendungsdaten auch die Qualitätsmerkmale explizit beschreiben bzw. verarbeiten können. Entsprechende Verfahren existieren bereits in verschiedenen Disziplinen für Artefakte auf jeder Abstraktionsstufe. Hunter und Goodchild [HG93] beschreiben beispielsweise eine Methode, die im Datenmanagement angewendet werden kann. Aus dem Forschungsgebiet der *Datenintegration* stammen Methoden, mit denen mehrere Datenbestände zu einem Referenzdatensatz konsolidiert werden. Dies schließt in vielen Fällen die Identifizierung und Eliminierung von Inkonsistenzen mit ein.
Die Modellierung von Unsicherheitsfaktoren beschränkt sich nicht auf die Daten, sondern auch auf nachgeordnete Artefakte. Die Idee, dass auch Musterartefakte und Konzepte nicht immer exakt (d.h. im Sinne der zweiwertigen Logik) beschrieben werden können wird durch deren Modellierung als *Fuzzy Sets* - so genannte unscharfe Mengen - geprägt (siehe z.B. Olson und Delen [OD08]). Entsprechende Inferenzmethoden für die Verarbeitung dieser Artefakte ergänzen das Repertoire an Techniken für die Transformation zwischen den analytischen Artefakten. Griethe und Schumann [GS06], sowie MacEachren et al. [MRH+05] geben einen Überblick über die entsprechenden Methoden aus der Informationsvisualisierung, die analog dazu Merkmale der Datenqualität visuell abbilden.
All diese Ansätze operieren auf der Ebene der Repräsentation der analytischen Artefakte. Die Varianten der Methoden, mit denen die Informationen über die Unsicherheit zwischen verschiedenen Abstraktionsstufen übersetzt werden, modellieren gleichzeitig auch die Propagation der Unsicherheitsfaktoren (siehe Abbildung 2.3). Belastbar werden Analysen aber nicht allein dadurch, dass man sicherstellt, dass die Daten allen Ansprüchen an eine korrekte Referenz genügen. Selbst wenn alle Unsicherheitsfaktoren in den Daten eliminiert werden könnten garantiert dies noch kein aussagekräftiges Ergebnis für die Analyse. Einen ebenso großen Unsicherheitsfaktor kann in der Praxis die Auswahl der Methoden und die Wahl ihrer Parameter darstellen.
Das Ergebnis der Analyse ist von der Qualität der Informationen ebenso abhängig wie von

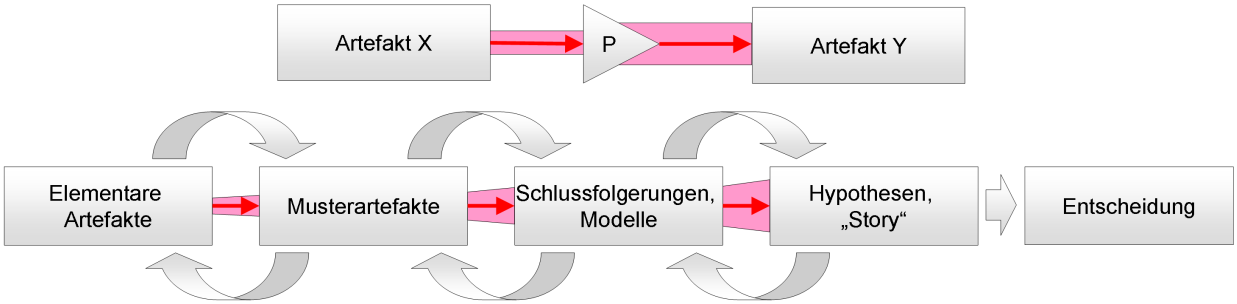


Abbildung 2.3: *Unsichere Information pflanzt sich innerhalb des analytischen Prozesses fort. Dabei können die Fehler in jedem Schritt P der Analyse propagiert werden. Wird die Entwicklung dieser Fehler nicht überwacht, kann nicht entschieden werden, ob die Ergebnisse überhaupt eine Aussagekraft besitzen. In verschiedenen Technologien gibt es Ansätze, die gleichzeitig mit den Artefakten, die sich auf die Fragestellung der Analyse beziehen, auch eine Beschreibung der Unsicherheit modellieren. Methoden, in denen zusätzlich auch diese Beschreibung verarbeitet wird, werden eingesetzt, um die Propagationsfehler zu kontrollieren und Fehler zu identifizieren. Aussagen über die Qualität der Prozesse selbst lassen sich alleine daraus jedoch nicht ableiten.*

den Verfahren, mit denen diese Informationen verarbeitet werden. Grundsätzlich muss dabei untersucht und sichergestellt werden, dass das Ergebnis der Analyse kein Artefakt einer Analysemethode oder der Art ihrer Anwendung ist (siehe Abbildung 2.4). In der Praxis ist es häufig möglich, in einer Analyse allein dadurch verschiedene Ergebnisse zu „gestalten“, in dem Wahl oder Parameter der Verfahren geändert werden. Ein bekanntes Beispiel ist die Normalverteilung der Ergebnisse von Intelligenztests (siehe Lewis [LEW57]). Diese Verteilung ist kein Artefakt der eigentlich zu messenden Daten, sondern ein Artefakt der Messung selbst. *Unter der Annahme*, dass Intelligenz in der Bevölkerung normalverteilt ist, werden die Tests so kalibriert, dass die Ergebnisse einer Normalverteilung entsprechen. Selbst wenn sie aus der Präsentation der Ergebnisse abgelesen werden könnte, ist die Normalverteilung dennoch kein gültiges Ergebnis der Analyse. In der Analyse sollte man erwarten können, dass auch die Entscheidungen über die Verfahren belastbar sind. Geschieht die Änderung dagegen willkürlich oder orientiert sich die Analyse gar ein einem postulierten Ergebnis, verliert das Ergebnis an Aussagekraft.
Datenqualität und -unsicherheit wird als eine der „Top 10 Research Challenges“ im Bereich Visual Analytics [KMS+08] genannt. Den Entscheidungen über die Auswahl von Verfahren und ihrer Steuerung kann jedoch für die Analyse die gleiche Relevanz eingeräumt werden, wie der Qualität der zugrundeliegenden Referenzartefakte. In dieser Arbeit soll ein Ansatz vorgestellt werden, mit dem ein Verfahren in der gleichen Weise bewertet werden kann, wie die Modelle, die es konstruiert. Die zentrale Forderung nach Transparenz wird damit nicht nur auf die Analyse der Daten bezogen, sondern auf die Analyse der Analyse selbst.
Innerhalb der Kette der analytischen Artefakte umfasst der Schwerpunkt dieser Arbeit im wesentlichen die elementaren Artefakte, die Musterartefakte und der Modelle und damit auch die Methoden, mit denen diese jeweils ineinander überführt werden können. Diese Methoden sind gerade deshalb für eine Analyse besonders interessant, weil gerade bei der Abstraktion von Daten die Informationsmenge häufig um einen bedeutenden Anteil reduziert werden muss.

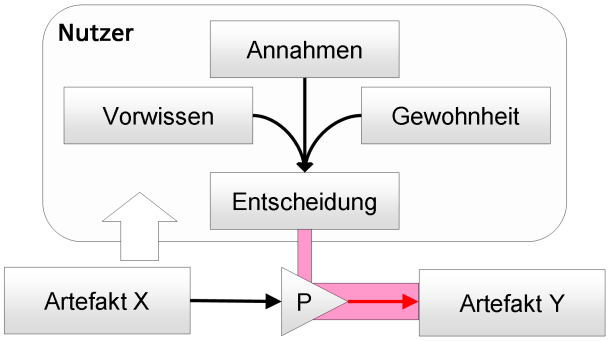


Abbildung 2.4: *Selbst wenn man davon ausgeht, dass keine Unsicherheit über die Korrektheit der analytischen Artefakte X herrscht, kann diese Unsicherheit durch die Entscheidungen des Nutzers bei der Analyse selbst entstehen. Sofern es keine gesicherte Regel dafür gibt, welche Verfahren wie auf welche Daten angewendet werden müssen, besteht immer die Möglichkeit, dass die Entscheidung des Analysten potentiell falsch ist. Da eine Vielzahl der Optionen des Nutzers mittelbar aus den Daten und Fragestellungen folgen, lassen sich nicht immer allgemeine Regeln angeben. Eine Herausforderung bei der Analyse besteht daher darin, auszuschließen, dass die Ergebnisse ein Artefakt des eingesetzten Verfahrens und seiner Parameter sind.*

Diese Reduktion ist erwünscht, da sie die Grundlage für die Beschreibung allgemeiner - nicht fallbezogener - Zusammenhänge darstellt. Allerdings muss man dafür in Kauf nehmen, dass durch die Reduktion ein Fehler in die Analyse eingebracht wird; und zwar selbst dann, wenn die zugrundeliegenden Daten korrekt sind. In allen im folgenden Abschnitt 2.3 vorgestellten automatischen Verfahren kann ein solcher Fehler quantitativ abgeschätzt werden (vgl. auch [FPSS96b]). Allerdings liefert ein Fehlermaß nur quantitative Informationen, jedoch keine Hinweise über die Qualität dieses Fehlers. Beispielsweise ist anhand des Fehlermaßes nicht zu unterscheiden, ob der Fehler sich in einem Hintergrundrauschen, in einem Ausreißer oder gar in einem verborgenen Muster manifestiert. Die Wahl eines Fehlermaßes determiniert die Trennung zwischen Muster und Rauschen.
In der Statistik unterscheidet man zwischen parameterfreien und parametrischen Modellen und Testverfahren [LW06]. Dabei wird jedoch nicht unterschieden, ob die Modelle überhaupt Parameter besitzen, sondern ob die Parameter ausschließlich aus den Daten berechnet werden (und damit erst im nachhinein bekannt sind) oder ob es Parameter gibt, die für Verfahren oder Tests vorher festgelegt werden müssen. Letztere sind die Freiheitsgrade für die Steuerung der Verfahren und die Gestaltung der erzeugten Modelle.
Über die belastbare Wahl eines oder mehrerer Verfahren, erhöht sich mit der Anzahl der Freiheitsgrade auch die Komplexität der Analyse für den Nutzer. In den folgenden Abschnitten wird deutlich, dass praktisch alle Methoden sowohl des Data-Mining als auch der Informationsvisualisierung durch den Nutzer über eine Reihe von Parametern gesteuert werden können. Dabei ist zu erwarten, dass die Wahl dieser Parameter einen Effekt auf das Ergebnis dieser Methode hat. Typische Beispiele für mögliche, wählbare Parameter in Data-Mining Methoden etwa sind:

- Vorgabe der Anzahl der Cluster bei bestimmten Clusteringverfahren (etwa *K-Means*)
- Vorgabe des Distanzmaßes zwischen Merkmalsvektoren bei Clusteringverfahren

- Gütefunktionen für die Heuristiken, die Modellparameter optimieren

Ein Beispiel für besonders kritische Freiheitsgrade bei Visualisierungstechniken betreffen die Wahl der (bei der Visualisierung praktisch immer beschränkten) Attribute eines Datensatzes, die gleichzeitig dargestellt und in Bezug gesetzt werden sollen. Die Informationen können damit immer nur ausschnittsweise betrachtet werden. In allen Fällen in den die relevanten Zusammenhänge mehr Merkmale umfassen, besteht die Gefahr, dass diese nicht gefunden bzw. nicht getestet werden. Jeder dieser Freiheitsgrade repräsentiert einen Parameter des Verfahrens oder des Modells, dessen Wert entweder

- abhängig von zusätzlichem Wissen über die Daten und ihre Eigenschaften,
- abhängig von Annahmen über die Daten und ihre Eigenschaften oder
- abhängig von einer eigenen Analyse

gesetzt werden muss. Problematisch im Sinne eine belastbaren Analyse ist, das einem Ergebnis nicht mehr anzusehen ist, ob die Parameter auf der Basis von Wissen oder auf der Basis von Annahmen gesetzt wurden.
Bei hinreichender Komplexität der Analyse kann es Parameter geben, die aus den Zielen des Anwendungsszenarios direkt abgeleitet werden könnten. Um die Wahl von Verfahren und Parametern zu begründen, müssen daher bei der Analyse entsprechende Qualitätskriterien entwickelt bzw. angewendet werden. In der Form von Gütefunktionen werden etwa Qualitätskriterien beschrieben, die die Heuristiken automatischer Data-Mining Verfahren steuern. Entsprechende Kriterien müssen aber auch angewendet werden, um ggf. die Entscheidungen des Nutzers bei der Analyse zu erklären.
Daraus ergibt sich jedoch die zentrale Fragestellung, wie man feststellt, dass die verwendeten Qualitätskriterien selbst den Anforderungen der Analyse genügen. Prinzipiell muss auch die Wahl der Kriterien so lange kritisch hinterfragt werden können, bis sie mittelbar oder unmittelbar aus den Zielen der Analyse abgeleitet werden können. Nach dieser Anforderung können die Entscheidungen innerhalb der Analyse als Hierarchie aufgefasst werden. Die Wahl eines Kriteriums bestimmt das Bezugssystem, innerhalb dessen untergeordnete Entscheidungen bewertet werden.
Im kleinsten Rahmen stellt sich zum Beispiel die Frage, welche Gütefunktion für ein Data-Mining verwendet werden soll, um in einem Datensatz Muster vom Rauschen abzugrenzen. Übergeordnet ist die Frage, ob das Verfahren prinzipiell in der Lage ist, die Muster in dem Datensatz zu identifizieren. Die Gütefunktion macht nur Aussagen über die Qualität der Muster im Kontext eines Verfahrens, jedoch keine Aussagen über die Qualität des Verfahrens selbst. Hierfür muss ein größerer Bezugsrahmen gefunden werden. Selbst wenn man schließlich sicherstellen kann, dass die bestmöglichen Verfahren und Parameter für die Analyse eines Datensatzes gefunden wurden, bleibt zum Beispiel die Frage, inwiefern die Analyse als Gesamtheit aller eingesetzten Verfahren auf andere Daten übertragen werden kann.

Die Untersuchung der Kriterien entspricht den Fragestellungen, die bei Evaluierungstudien für Verfahren, „Best-Practices“ oder auch für Methodiken und Systeme behandelt werden.

Daher muss an dieser Stelle betont werden, dass der Fokus dieser Arbeit *nicht* eine Evaluierung im Sinne einer Vergleichstudie für verschiedene Analyseerfahren ist. Der Fokus liegt vielmehr auf einer Verbesserung der Möglichkeiten für die Evaluierung während der Analyse selbst. Eine Evaluierung ist schließlich auch ein Teilschritt in Prozessmodellen für die Datenanalyse u.a. von Chapman et al und Fayyad et al. [CCK+00, FPSS96b], und der Ausgangspunkt für einen Eingriff des Nutzers. In dieser Arbeit sollen Methoden entwickelt werden, mit denen die Evaluierung während der Analyse transparenter gestaltet werden kann.
Der Unterschied besteht darin, dass in einer Evalierungsstudie die Wahl des Untersuchungsgegenstands (z.B. eine Klasse von Analyseverfahren) und der Kriterien einen klar umrissenen Kontext definiert. Innerhalb dieses Kontexts können die Ergebnisse der Studie verglichen werden. McGrath [McG95] unterscheidet verschiedene Evaluierungsstrategien für Studien aus den Sozialwissenschaften und stellt fest, dass jede Strategie einen „Trade-Off“ darstellt zwischen Verallgemeinerbarkeit, Präzision, und Wirklichkeitstreue der Ergebnisse. Alle drei Anforderungen können in einer Evaluierung also prinzipiell nicht in gleich hohen Standard erfüllt werden. Nach Plaisant [Pla04], sowie Zuk und Carpendale [ZC07] lassen sich diese Aussagen auf Evaluierungsstudien in der Informationsvisualisierung übertragen.
In der Praxis können und sollten Ergebnisse von Evaluierungsstudien bei der Wahl von Verfahren genutzt werden. Allerdings lässt sich daraus nie eine absolute Sicherheit bei der Wahl der Verfahren ableiten. Gründe dafür sind beispielsweise darin zu suchen, dass ein Verfahren in der Praxis in einem Kontext verwendet wird, der nicht dem entsprechen muss, für den es evaluiert wurde bzw. für den es ursprünglich entwickelt wurde. Dies kann in unterschiedlichen Fällen geschehen:
Wird ein Analyseverfahren übertragen auf ein neues Anwendungsgebiet, so muss auch das Verfahren mittelbar oder unmittelbar an die Kriterien des Analyseziels angepasst werden. Beispielsweise sind die Anforderungen an die Spezifität eines Prädiktors - der Anteil der korrekt als negativ klassifizierten Fälle - in der Krebsforschung anders als im Customer Relationship Management. Dieses Beispiel zeigt auch, dass eigentlich technische Parameter eine inhaltliche Bedeutung besitzen können, die so groß ist, dass der Entscheidungsträger (d.h. der „Endanwender“) selbst die Verantwortung dafür übernehmen muss. Je indirekter diese technischen Parameter mit dem Analyseziel verbunden sind, desto schwieriger ist diese Anpassung von vornherein einzuschätzen.
Ein anderes Problem sind Anwendungsszenarien, in denen dynamische Daten ausgewertet werden, die sich dadurch auszeichnen, dass die Datenmuster, die die relevanten Informationen enthalten, sich permanent ändern. So findet zum Beispiel etwa bei der Bekämpfung der Onlinekriminalität eine Koevolution zwischen Kriminellen und Kriminalisten statt, die dazu führt, dass die Verhaltens- und Transaktionsmuster, nach denen gefahndet werden muss, stetig weiterentwickeln. Die Anwendbarkeit der Analyse auf die aktuellen Daten muss im Prinzip ständig einer (Neu-)Bewertung unterzogen werden.
In Analysen werden Analyseverfahren vielfältig miteinander kombiniert. Durch die Rekombination ergibt sich einerseits ein großes Spektrum an Möglichkeiten. Andererseits ist nicht auszuschließen, dass die Verfahren so wechselwirken, dass das Ergebnis der Analyse in unvorhergesehener Weise beeinflusst wird. Eine solche Wechselwirkung kann zum Beispiel zwischen der Auswahl der Daten und ihrer Weiterverarbeitung bestehen.
Zuletzt besteht die Herausforderung eines Analysten darin, einen Prozess zu gestalten, ohne jedoch *alle* Informationen zu kennen, nach denen die belastbare Wahl von Verfahren

oder Parametern erfolgen könnte. Westphal und Blaxton [WB98, Seite57] beschreiben dies als „Kochen ohne Rezept“. Die erste Entscheidung betrifft die für die Analyse verwendeten Daten. Im Extremfall kann sie nur aus der Zielvorgabe und dem Vorwissen über den Gegenstand der Analyse abgeleitet werden. Fast alle folgenden Entscheidungen basieren aber zu einem bedeutenden Teil auch auf den Informationen, die erst während der Analyse in den Daten gefunden wurden. Der gesamte Prozess der Analyse für die Fragestellungen eines Anwendungsgebiets setzt sich bis zum kleinsten Teilprozess fort. Zusätzlich zur Belastbarkeit der Kriterien einer Entscheidung kommt mithin die Belastbarkeit der Informationen, die ihr zugrundeliegen.
Da in einer Analyse natürlich mindestens das Wissen nicht vorliegt, das erst aus der Analyse geschöpft werden soll, folgt aus diesen Überlegungen, dass Entscheidungen auf der Basis nicht gesicherten Wissens zunächst in Kauf genommen werden müssen. Um die Analyse belastbar zu machen, kann man fordern, dass jede dieser Entscheidungen erkannt, analysiert und ggf. revidiert wird.
Die Bezugsgrundlage für die Bewertung der Analyse sind genau jene analytischen Artefakte, deren Korrektheit im Rahmen der Analyse nicht angezweifelt werden. Dies schließt wohlgemerkt nicht aus, dass sie im Rahmen einer anderen Analyse sehr wohl belastet werden könnten. Die (Roh-)daten sind gerade jene Referenzinformationen für eine Bewertung auf der Ebene der elementaren Artefakte. Die Bewertung über „Ground-truth“-Methoden entspricht gerade dem Vergleich auf dieser Abstraktionsebene.
Entsprechende Referenzen existieren jedoch - ob explizit oder implizit - auf allen Abstraktionsebenen. Da es sich beispielsweise bei Mustern und Begriffen gleichermaßen um Mengen handelt, ist es möglich, neu gefundene und bereits bekannte Information auf dieser Ebene zu vergleichen. Auf der höchsten Abstraktionsstufe wäre es möglich, Annahmen und Hypothesen gegen bestehendes Wissen zu prüfen. Allgemein entspricht eine Bewertung der Analyse entweder einer Prüfung eines „Testartefakts“ an einem als sicher erachteten Referenzartefakt oder einer Prüfung von Artefakten, die als gleich (un)sicher erachtet werden, zum Beispiel in der Form eines Plausibilitätstests.

2.1.2 Strategien der Analyse

Die Bekräftigung oder Widerlegung von Hypothesen anhand geeigneter Referenzen, ist Gegenstand der *konfirmativen Datenanalyse* [FPSS96b]. Wenn eine Hypothese bereits bekanntem Wissen nicht offensichtlich widerspricht oder entspricht, besteht das Ziel der Analyse darin, die Artefakte so zu transformieren, dass ein unmittelbarer Vergleich mit anderen Referenzartefakten möglich wird. Welche Artefakte umgeformt werden, ist dabei abhängig von der gewählten Methodik. Zu den ältesten Methoden der konfirmativen Datenanalyse gehören die Hypothesentests der mathematischen Statistik [LW06]. Abhängig von den Anforderungen an den Test, wird aus der Hypothese ein Entscheidungskriterium konstruiert, das auf statistische Kenngrößen von Testdaten angewendet werden kann.
Einen anderen Ansatz für die konfirmative Datenanalyse bietet die *Simulation*. Wenn eine Hypothese Voraussagen über die Daten zulässt, dann lässt sich die Hypothese durch den Vergleich dieser Voraussagen mit realen Daten auf der Ebene elementarer Artefakte mindestens qualitativ beschreiben.

Nicht Teil dieser Arbeit ist eine konfirmative Datenanalyse mittels logischer Inferenzmethoden. Dabei handelt es sich um eine Transformation auf Artefakten höherer und höchster Ordnung. Liegt Wissen über die Anwendungsdomäne in der Form einer Wissensbasis explizit vor, und ist die Hypothese eine innerhalb der Domäne beschreibbare Aussage, bestünde die Möglichkeit, die Hypothese durch logisches Schlussfolgerungen zu testen und ggf. in die Wissenbasis zu integrieren.
Das Gegenstück zur konfirmativen ist die *explorative Datenanalyse* [FPSS96b, TC05]. Sie wurde in den siebziger Jahren von Tukey [Tuk77] etabliert. Hypothesen über einen Datenbestand sind dabei nicht der Ausgangspunkt der explorativen Datenanalyse, sondern ihr Ziel. Gesucht werden neue, nicht-triviale Aussagen, die den Aufwand rechtfertigen, sie zu prüfen. Konfirmative Datenanalyse und explorative Datenanalyse lassen sich nicht als zwei grundsätzlich unterschiedliche Methodiken der Analyse begreifen, die frei und unabhängig voneinander gewählt werden können. Shneiderman legt in [Shn02] die Schwächen beider Strategien dar und betont, dass beide Strategien für die Datenanalyse die Schwächen der jeweils anderen kompensieren könnten.
Hypothesentests erzwingen, dass die zu überprüfenden Aussagen ausreichend präzise artikuliert und Ergebnisse auch interpretiert werden können. Die Präzision erkauft man sich aber dadurch, dass „Seiteneffekte“ - wie zum Beispiel auch der Auffinden unerwarteter Muster - methodisch eliminiert werden. Bei der explorativen Analyse sind diese Seiteneffekte gerade erwünscht. Ihre Schwäche liegt allerdings darin, dass ihre Methoden nicht notwendig eine Garantie bezüglich der Aussagekraft, Relevanz oder der Interpretierbarkeit der Ergebnisse bieten.
Für beide Strategien beschreibt Shneiderman *(ebd.)* jeweils das Repertoire der üblicherweise verwendeten Verfahren. Er zählt Data-Mining Techniken zum „traditionellen“ Repertoire für die explorativen Datenanalyse und statistische Testverfahren zum Repertoire für Hypothesentests. Allerdings beruht nach Shneiderman die Zuordnung der Verfahren zu einer bestimmten Methodik nicht auf Eigenschaften der Verfahren selbst, sondern begründet diese durch die traditionellen Methodiken der Forschungsgebiete, in denen die Verfahren entwickelt wurden. Shneiderman schlägt deshalb eine Synthese beider Methodiken vor, die durch einen Wechsel und Austausch zwischen präziser Beschreibung auf der einen und flexibler Exploration auf der anderen Seite gekennzeichnet ist.
Methoden der Informationsvisualisierung werden sowohl für die konfirmative als auch für die explorative Datenanalyse eingesetzt (Keim et al. [KMSZ06]). In der konfirmativen Datenanalyse ermöglichen sie den visuellen Vergleich zwischen Entscheidungskriterien, Referenzdaten und/oder Testdaten oder aber auch den visuellen Abgleich zwischen mehreren Testgrößen. Hypothesentests auf der Basis rein statistischer Verfahren erlauben wegen der notwendigen Aggregation nur quantitative Aussagen über die Beziehung zwischen Testdaten und Testkriterien. Unter der Vorraussetzung, dass eine Hypothese auf der Basis zahlreicher Referenzdaten überprüft werden muss, kann eine Visualisierung dabei helfen, Referenzdaten und Testdaten im gleichen Zusammenhang darzustellen und damit auch qualitative Aussagen zu erhalten. Ein möglicher Ansatz wird dazu wird im Rahmen dieses Konzepts beschrieben.
In der explorativen Datenanalyse dienen interaktive Visualisierungstechniken ebenso wie auch Data-Mining Verfahren der Identifikation von Mustern. Shneiderman und andere Autoren [WB98, Shn02, Kei02, TC05, vW05] legen jedoch dar, dass die Methoden beider Forschungsgebiete unterschiedliche Stärken besitzen. Die Stärken der automatischen Verfahren

(d.h. der Data-Mining Verfahren) beruhen auf den Vorteilen, die die maschinelle Verarbeitung bietet. Technisch ermöglicht dies die detaillierte Untersuchung großer Datenmengen unter Berücksichtigung zahlreicher abhängiger und unabhängiger Dimensionen.
Ebenso wichtig ist, dass ein automatischer Prozess auch einen methodischen Vorteil bietet. Im Sinne einer kritischen Untersuchung stellt der Prozess gewissermaßen ein wiederholbares Experiment dar, dessen Parameter präzise kontrolliert werden können.
Die Schwäche der automatischen Verfahren bestehen darin, dass die Klassen der Muster, die überhaupt erkannt und beschrieben werden können, von vornherein durch die Entwickler der Verfahren festgelegt werden. Diese Menge kann man vergrößern, indem man die Mächtigkeit der Beschreibungssprache mit der Anzahl ihrer Freiheitsgrade erhöht. Allerdings erhöht sich dabei im Allgemeinen sowohl die Komplexität der Berechnung und auch die Komplexität der Interpretation der Ergebnisse - ohne dass diese Beschränktheit prinzipiell aufgehoben werden könnte.
Die Stärken der visuell-interaktiven Verfahren beruhen auf der Nutzung der spezifischen Fähigkeiten des Menschen zum kreativen und flexiblen Problemlösen, seiner visuellen Wahrnehmung, seines Wissens und auch der Fähigkeit Begriffe zu entwickeln und zu benennen, die die Grundlage für die Artikulation neuen Wissens darstellen [SP04, Seite 79]. Visualisierungen ermöglichen die flexible und effiziente Erkennung von Musterartefakten durch den Menschen. Dadurch wird es möglich, jene automatischen Verfahren zu bestimmen, welche auf die Daten angewendet werden können. Darüber hinaus wird es möglich, auch solche Muster zu erkennen für die (noch) kein Extraktionsverfahren existiert.
Die Grenzen in der interaktiven Visualisierung liegen in der naturgegebenen Beschränkung der zwei Dimensionen des Bildschirms und die Beschränkung der Aufmerksamkeit des Menschen. Die Menge an Daten und mehr noch die Anzahl der Dimensionen, die gleichzeitig dargestellt und wahrgenommen werden können, ist begrenzt. Beispielsweise Ward [War04a] erörtert Methoden um mit dieser Beschränkung umzugehen, die an verschiedenen Stufen des Visualisierungsprozesses ansetzen. Auf die Stärken und Schwächen von Methoden beider Technologien und die darauf folgenden Konsequenzen wird in den folgenden beiden Abschnitten 2.2 und 2.4 im Detail eingegangen.

2.1.3 Schwerpunkte dieser Arbeit

Die Integration von interaktiven Visualisierungstechniken in die Datenanalyse soll dabei helfen, die dahinterliegenden Prozesse transparent zu machen und dem Nutzer für den analytischen Diskurs zu exponieren. Dies ist die zugrundeliegende Motivation für die Forschungsgebiete *Visual Analytics* bzw. (das oft synonym verwendete) *Visual Data Mining*. Thomas & Cook [TC05] definieren Visual Analytics als "Wissenschaft des analytischen Schließens, unterstützt durch interaktive visuelle Schnittstellen".
In dem Forschungsgebiet wird nicht nach Methoden gesucht, mit denen die Prozesse interaktiv durchgeführt werden können, die besser automatisch durchgeführt werden sollen (oder umgekehrt). Zentrale Motivation ist vielmehr die Einsicht, dass interaktive und automatischen Methoden unterschiedliche Schwächen und Stärken haben, und dass für eine belastbare Analyse die Stärken nur einer Technologie nicht ausreichen. Die Kombination von Techni-

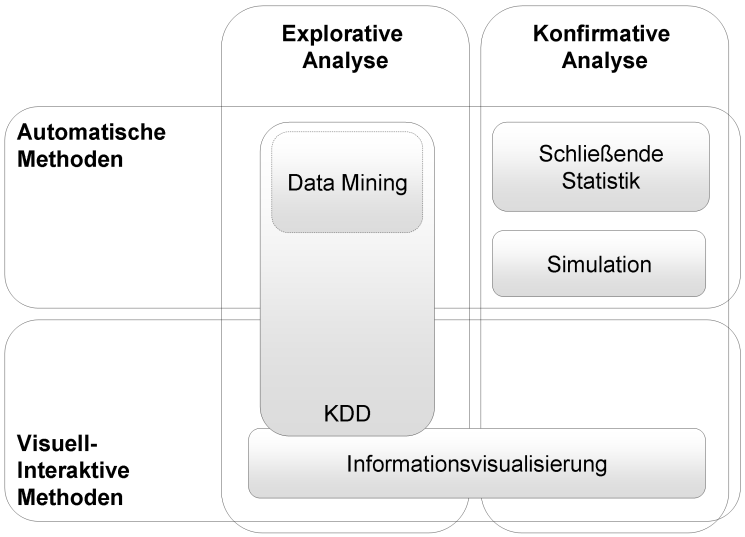


Abbildung 2.5: *Der Schwerpunkt dieser Arbeit umfasst die hier gegebenen Forschungsgebiete unter zwei wesentlichen Aspekten. Der erste Aspekt ist die Kopplung zwischen automatischen und visuell-interaktiven Methoden; der zweite Aspekt ist die Verbindung zwischen explorativer und konfirmativer Datenanalyse.*

ken für die automatische und interaktive Datenanalyse soll dabei helfen, die Schwächen beider Technologien zu komponsieren. Dies ist der Ausgangspunkt für praktisch alle Ansätze in diesen Forschungsgebieten. In den folgenden Abschnitten 2.2, 2.3 und 2.4 sollen die Forschungsgebiete des Knowledge-Discovery, des Data-Mining und der Informationsvisualisierung untersucht werden, um

- die Stärken und Schwächen der Methoden in beiden Technologien zu beschreiben,
- Ansatzpunkte für mögliche Varianten von Kopplungen zwischen beiden Technologien zu identifizieren.

In der aktuellen Forschung präsentiert das Gebiet Visual Analytics sehr heterogen. Die verschiedenen Ansätze lassen sich grob klassifizieren nach

- Anwendungsgebieten,
- den jeweils miteinander gekoppelten Verfahren,
- der Art der Kopplung.

Diese Arbeit ist nicht ausgerichtet auf die Beschreibung von Ansätzen für bestimmte Anwendungsgebiete oder bestimmte Verfahren. Stattdessen liegt der Fokus auf der Beschreibung möglicher Kopplungen. Im Abschnitt 2.5 werden existierende Modelle vorgestellt, mit denen verschiedene Ansätze nach diesem Kriterium eingeordnet werden. Im Konzept dieser Arbeit werden Ansätze für eine Kopplung vorgestellt, wobei zwei Aspekte im Vordergrund stehen (siehe Abbildung 2.5).

Der erste Aspekt ist die Aufteilung des Transformationsprozess von Daten über Mustern hin zu Modellen, derart, dass der Mensch die Teilaufgabe der Mustererkennung übernimmt, und der Maschine die Teilaufgabe der Musterbeschreibung.
Der zweite Aspekt, ist die Kombination explorativer und konfirmativer Datenanalyse. Im Konzept soll untersucht werden, wie über die Kopplung automatischer und interaktiver Methoden der Datenanalyse eine Integration von explorativer und konfirmativer Datenanalyse erreicht werden kann. Ziel dieser Integration ist nicht alleine eine unmittelbare Bestätigung der durch explorative Analyse gefundenen Modelle. Ziel ist auch eine transparente Bewertung der dafür eingesetzten Verfahren.
Diese beiden Aspekte entsprechen den beiden ersten Hauptzielen dieser Arbeit. Das dritte Ziel wird durch eine Synthese dieser beiden Aspekte umgesetzt, innerhalb derer der Mensch nicht allein die Aufgabe für die Mustererkennung in der explorativen Datenanalyse übernimmt, sondern allgemeiner den visuellen Abgleich zwischen Referenzartefakten und den Artefakten, die in der Analyse konstruiert werden.

2.2 Knowledge Discovery in Databases (KDD)

In der Praxis vieler Anwendungsgebiete besteht die Herausforderung der Analyse darin, dass die Kluft zwischen der Menge potentiell relevanter, verfügbarer Informationen und der Menge der Informationen, die vom Menschen erfasst und verarbeitet werden können, immer größer wird. Gründe dafür sind unter anderem sicher ein hoher Grad an Automatisierung beim Sammeln von Daten und natürlich die technische Möglichkeit, überhaupt eine große Menge an Daten zu speichern und abzurufen. Diese Menge muss bei der Analyse auf die Information reduziert werden, die für eine gegebene Fragestellung relevant ist. Daraus folgt auch, dass in mindestens einem Teilschritt der Analyse eine Reduktion bezüglich der Datenmenge und/oder Datenkomplexität vorgenommen werden muss.
Das Forschungsgebiet Knowledge Discovery in Databases (KDD) stellt einen Rahmen vor, in dem verschiedene Techniken mit dem Ziel verbunden werden, die Informationen in den Datenbeständen so aufzubereiten, dass daraus Wissen geschöpft werden kann. Fayyad et al. [FPSS96b, FPSS96a] definieren das Ziel des Knowledge-Discovery als *nicht trivialen Prozess für die Suche nach validen, neuen, potentiell relevanten und nutzbaren Mustern in Datenbeständen.* Häufig wird der Begriff *Data-Mining* so besetzt, dass er sich mit *Knowledge Discovery in Databases* vollständig deckt. Innerhalb dieses Rahmens stellt Data-Mining jedoch nur einen, wenn auch zentralen, Teilschritt dar. Im Namen deutlich betont wird das umfassendere Ziel beider Technologien: Zentraler Anspruch des KDD-Prozess ist die Schöpfung und Nutzung von *Wissen*, während die Ziele, die im Allgemeinen für den Data-Mining Prozess reklamiert werden - die Entdeckung und Beschreibung von Mustern - als Unterziel des ganzen Prozesses eingeordnet werden.
Die Suche nach Mustern beschreibt dabei nur das allgemeine Ziel für den Data-Mining Schritts. Fayyad et al. unterscheiden *(ebd.)* feiner zwei Typen von Zielen. Das erste dieser Ziele ist die *Vorhersage* bzw. *Prädiktion*, mit der es möglich sein soll aus den gegebenen Daten zukünftige Werte abzuleiten. In vielen Fällen deutet der Begriff „Vorhersage“ auf einen zeitlichen Zusammenhang zwischen bekannten und abhängigen Variablen eines Datensatzes hin. Für die Präsentation der Ergebnisse könnte es bisweilen notwendig sein, zwischen zeitlichen und anderen Zusammenhängen zu unterscheiden. Jedoch umfaßt dieses Ziel im Rahmen dieser Arbeit die Identifizierung *beliebiger* funktionaler Zusammenhänge zwischen Attributen einer Entität.
Das zweite dieser Ziele ist die *Beschreibung* der Daten. Im Sinne der Definition des KDD-Prozesses wird dabei gefordert, dass diese Beschreibung durch den Menschen lesbar ist.
Das Modell des KDD-Prozess beschreibt eine semi-automatische Datenverarbeitungspipeline (siehe Abbildung 2.6). Es definiert einerseits die Abfolge aller Teilschritte, die alle für die Analyse notwendig sind, andererseits exponiert sie spezifische Ansatzpunkte für die Bewertung und die Eingriffe durch den Menschen. Zur Analyse gehören dabei sowohl automatische wie nicht-automatische Teilprozesse.

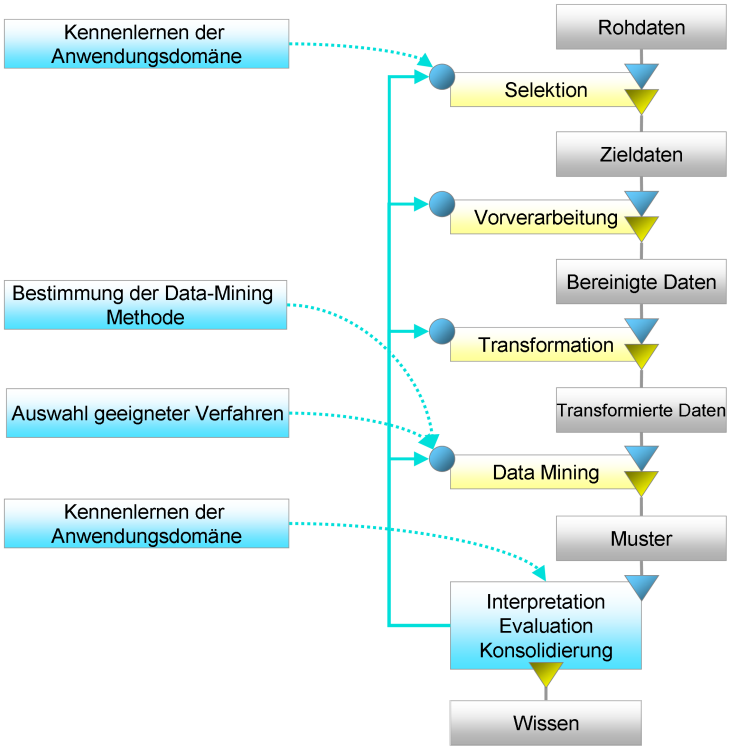


Abbildung 2.6: *Das Modell des Knowledge-Discovery-Prozesses (nach [FPSS96b]) Es handelt sich dabei um eine Datenverarbeitungspipeline, in der jeder Schritt durch Interaktion auf der Grundlage von Zwischen- oder Endergebnissen modifiziert werden kann. Zusätzlich zu den technischen Abläufen enthält das Modell jedoch auch nicht-technische Teilschritte. Der Teilschritt „Kennenlernen der Anwendungsdomäne“ beeinflusst alle Schritte der Analyse. Jede Iteration kann zwischen zwei beliebigen Teilschritten dieses Prozesses stattfinden; jeder Schritt kann also Ergebnisse liefern, auf deren Basis vorangegangene Schritte modifiziert werden können.*

Im Einzelnen sind die Schritte des Knowledge-Discovery:

1. *Kennenlernen der Anwendungsdomäne:* Das Hintergrundwissen aus der Anwendungsdomäne ist potentiell relevant für die Analyse; es ist notwendig für die Beurteilung der Relevanz von Daten oder Zwischenergebnissen, für die Wahl der Referenzartefakte, gegen die die Analyse selbst geprüft werden kann, und um die Fragestellung der Anwendungsdomäne korrekt in die der Analyse übersetzen zu können.

2. *Selektion der Daten:* Abhängig von ihrer Relevanz werden bestimmte Attribute oder Objekte für die Analyse ausgesucht. Dies ist bereits die erste Entscheidung, mit der das Ergebnis der Analyse signifikant beeinflusst werden kann.

3. *Datenvorverarbeitung:* Die Vorverarbeitung umfaßt Schritte wie die Datenintegration, Datenbereinigung, die Bewertung der Datenqualität und ggf. die Bestimmung von Kennzahlen für die Steuerung der folgenden Verfahren.

4. *Datentransformation und -reduktion:* In diesen Teilschritten können die Menge oder die Komplexität der Daten reduziert werden. Ziel dieser Prozesse ist die Anpassung

der Daten an die Anforderungen der danach genutzten Data-Mining-Verfahren, wobei der Informationsgehalt der Daten so gut wie möglich erhalten bleiben soll.

5. *Bestimmung der Data-Mining Aufgabe.* „Aufgaben" im Data-Mining im Sinne der Taxonomie von Han und Kamber [HK00] (siehe Abschnitt 2.3.2) sind beispielsweise Clustering oder Klassifikation. In diesem Schritt wird die Aufgabe abhängig von Zielen der Analyse identifiziert.

6. *Auswahl eines Data-Mining-Verfahrens.* Nachdem die Aufgabe definiert wurde, müssen in Abhängigkeit von Aufgaben und Daten geeignete Modelle und Verfahren identifiziert werden. Beispiele dafür werden in Abschnitt 2.3.5 behandelt.

7. *Data-Mining.* Anwendung des gewählten Verfahrens auf die transformierten Daten. Ergebnis dieses Schritts sind identifizierte Muster oder Modelle, die die Beziehungen in den Daten formal beschreiben.

8. *Interpretation.* Die im Data-Mining gefundenen Informationen sind nur Zwischenergebnisse, die zunächst interpretiert und evaluiert werden müssen. Die Ergebnisse der Analyse müssen dabei im Kontext der Anwendungsdomäne neu beschrieben werden.

9. *Nutzung des Wissens.*

Auch wenn der Data-Mining Prozess innerhalb des Modells als einer von mehreren Schritten beschrieben wird, steht dennoch das Data-Mining an exponierter Stelle. In diesem Schritt findet die Suche nach einer Beziehung zwischen elementaren Artefakten und Musterartefakten statt. Die Teilprozesse von der Selektion der Datenbasis bis zur Datentransformation können daher als vorbereitende Schritte interpretiert werden. Der folgende Abschnitt 2.3 setzt sich dediziert mit den Techniken des Data-Mining und Machine-Learning auseinander. Die Schritte nach dem Data-Mining sind gerade die Schritte, in denen wiederum eine Beziehung zwischen Musterartefakten und den analytischen Artefakten höherer und höchster Ordnung hergestellt wird. Nach der Interpretation und Evaluation der gefundenen Muster werden diese in der Wissensbasis der Anwender konsolidiert. Bemerkenswert dabei ist, dass das Modell des KDD-Prozess implizit keine *technische* Unterstützung für diese Schritte vorsieht, sieht man von einer Visualisierung der Ergebnisse ab. Die Hauptrolle der Visualisierung besteht jedoch darin, dass sie die Bearbeitung durch den Menschen durch die Exponierung der Ergebnisse überhaupt erst möglich macht. Sie wird als Präsentationsmedium für (Zwischen-) Ergebnisse betrachtet; für die Identifikation der Muster selbst sind Visualisierungsmethoden in diesem Modell nicht vorgesehen.

2.2.1 Bewertung und Einordnung des KDD-Prozess

Das Modell für den KDD-Prozess schreibt nur allgemein vor, mit welchen Techniken die einzelnen Schritte der Datenverarbeitungspipeline abgearbeitet werden. Der Prozess enthält dabei sowohl automatischen Komponenten (Teilprozesse, die ohne Eingriff des Menschen ablaufen), interaktive Komponenten (Teilprozesse, die durch den Nutzer gesteuert werden) und manuelle Komponenten (Teilprozesse bzw. Entscheidungen, die der Nutzer ohne technische

Unterstützung durchführen kann oder muss). Der KDD-Prozess fasst also sowohl automatische Prozesse als auch Entscheidungen des Nutzers in allgemeiner Form zusammen.
Die Entscheidungen des Nutzers können dabei sehr unterschiedlich stark strukturiert sein. Eine interaktive Komponente gibt durch ihre steuerbaren Parameter ein klar definiertes Schema für die möglichen Entscheidungsoptionen vor. Ein Prozess kann aber so unstrukturiert sein, wie etwa die Abstimmung von Methodiken, Datensätzen und Verfahren, für die kein vorgefertigtes Schema zur Verfügung steht. Der KDD-Prozess exponiert aber auch gerade die Entscheidungen und Eingriffe des Menschen als Teil der Analyse.
Die Einbettung des Data-Mining in einen größeren Rahmen, in dem sichergestellt wird, dass die Ergebnisse des Data-Mining valide und relevant sind, ist der Grund für diese Exposition. Fayyad et al. betonen damit insbesondere das Vorwissen, das in die Analyse investiert werden muss, wenn die Ergebnisse aussagekräftig sein sollen. Der KDD-Prozess wird dabei von der willkürlichen bzw. unsystematischen Anwendung eines Data-Mining Verfahrens auf beliebige Rohdaten ausdrücklich unterschieden. Die Ergebnisse der nicht-deterministischen Teilschritte 1., 5., 6., 8. und 9. sind Entscheidungen über die Methode, Verfahren, Parameter und Ergebnisse der Analyse. Das Modell gibt nicht im Detail vor, wie diese Entscheidungen getroffen werden sollen, sondern vielmehr, dass diese Teilschritte überhaupt als zu belastende Entscheidungen wahrgenommen werden. Hier wird die jeder Analyse zugrundeliegende Problematik deutlich; dass Wissen eine Voraussetzung dafür ist, Wissen aus der Analyse zu schöpfen.
Die Idee, ein Data-Mining Verfahren nicht willkürlich und unsystematisch auf beliebigen Daten durchzuführen entspricht methodischen Notwendigkeiten. Das Referenzmodell, das im *CRISP-DM* („Cross-Industry-Standard-Process for Data-Mining“) [CCK+00] vorgeschlagen wird, ist ein Industriestandard, der im Wesentlichen die gleichen Schritte zusammenfasst, die auch im KDD-Prozess definiert sind. Das Referenzmodell zeichnet die Abbildung eines bestimmten Anwendungsproblems auf einen KDD-Prozess und die Entscheidungen für oder wider bestimmte Verfahren in größerem Detail nach. Es stellt jedoch in der vorliegenden Version keine Theorie dar, nach der diese Entscheidung getroffen und erklärt werden können.
Dies zeigt auch, dass im KDD-Prozess zielgemäß Wissen über die in den Daten verborgenen Inhalte geschöpft werden soll - gleichzeitig wird bei der Konstruktion der Analyse aber auch Verfahrenswissen darüber gewonnen, *wie* diese Inhalte aus den Daten ableitbar sind. Auch wenn innerhalb der Anwendungsdomäne unter Umständen nur der erste Aspekt relevant ist, sind die beiden Ebenen dieser Wissenschöpfung eng miteinander verbunden und werden ggf. auch gemeinsam belastet. Im Konzept wird untersucht, welche Rolle Techniken der Informationsvisualisierung in diesem Prozess spielen können - über die Darstellung von (Zwischen-)Ergebnissen hinaus, und zwar sowohl für die Suche nach Mustern in den Daten, als auch als Feedback für die Entscheidungen des Analysten.
Das Modell für den Knowledge-Discovery-Prozess soll im Konzept dieser Arbeit als Referenz dienen für die Lokalisierung verschiedener Techniken und Prozesse und insbesondere für die Lokalisierung der Eingriffe des Nutzers. Jedoch deckt in der Darstellung von Fayyad et al. der Knowledge-Discovery-Prozess einige Aspekte *nicht* oder nur wenig detailliert ab, die für diese Arbeit relevant sind:
Ein von Fayyad et al. [FPSS96b] nur grob skizzierter Aspekt ist die Möglichkeit zur konfirmativen Analyse. Praktisch alle Teilschritte des Knowledge-Discovery - mit Ausnahme des Data-Mining und seiner Ergebnisse - lassen sich im Wesentlichen jedoch auf die kon-

firmative Analyse gegebener Hypothesen übertragen. Die Tatsache, dass eine gegebene Hypothese einen engeren Fokus etwa bei der Wahl der zu betrachtenden Daten und Attribute erlaubt, ist nur ein inhaltlicher, jedoch kein prinzipieller Unterschied. Nach Fayyad et al. *(ebd.)* ist die Verifikation von Hypothesen neben der Exploration eines der beiden Ziele des Knowledge-Discovery. Der Data-Mining-Schritt kann im Rahmen des Knowledge-Discovery als Teilprozess betrachtet werden, der charakteristisch ist für die explorative Analyse, wie auch Shneiderman [Shn02] bemerkt.
Wie oben erwähnt, sind Beschreibung und Vorhersage im Knowledge-Discovery die Teilziele der explorativen Datenanalyse. Alle Techniken des Data-Mining sind dementsprechend *deskriptive* oder *prädiktive* Verfahren, wobei viele Verfahren auch beide Teilziele erfüllen. Der Definition des Knowledge-Discovery nach gilt die Forderung, dass die Ergebnisse für den Menschen „verstehbar“ bzw. „lesbar“ sein müssen. Streng genommen schränkt dies das Repertoire einsetzbarer Verfahren entsprechend ein. Aus folgenden Gründen wird diese Einschränkung jedoch nicht vorgenommen:

1. Mit einer validen Vorhersage erhielte man, bezogen auf die Fragestellung der Anwendung, die Fähigkeit, eine Entscheidung anhand der verfügbaren Daten zu treffen. Mit einer validen Beschreibung erhält man Informationen über die Natur der mit den Daten repräsentierten Entitäten und ihrer Beziehungen. Fayyad et al. schreiben selbst, dass es in erster Linie vom Ziel der Analyse abhängt, welches der beiden Teilziele eher verfolgt werden soll.

2. Die Ergebnisse prädiktiver Verfahren sind Vorhersagen über die Daten. Im Rahmen einer konfirmativen Analyse lassen sich Vorhersagen intuitiver und direkter an Referenzdaten testen, als es mit rein deskriptiven Verfahren möglich wäre. Diese Möglichkeit ist einer der wichtigsten Voraussetzungen des Konzepts dieser Arbeit. Die Anwendung der Ideen dieses Konzepts auf rein deskriptive Verfahren wird im Ausblick diskutiert.

2.3 Automatische Verfahren - Data-Mining

Der Schwerpunkt dieses Abschnitts liegt in der Vorstellung des Repertoires automatischer Verfahren für die explorative Datenanalyse aus dem Data-Mining und Machine-Learning. Eine grundsätzliche Unterscheidung oder gar Trennung zwischen Verfahren des Data-Mining und denen des Machine-Learning wird dabei bewusst nicht gemacht. Traditionell wurden Techniken des Machine-Learning weniger mit Fokus auf der Wissensgewinnung und Analyse, denn auf die Optimierung (bzw. das „Lernen“) automatischer Prozesse (z.B. Robotik, Bild- und Texterkennung) entwickelt. Die gleichen Verfahren werden jedoch zunehmend auch im Rahmen der Datenanalyse eingesetzt. Im Rahmen des hier vorgestellten Konzepts ist eine Unterscheidung nicht notwendig.
Generell sollen in dieser Beschreibung die Gemeinsamkeiten zwischen einzelnen Verfahren stärker in den Vordergrund gerückt werden, als deren Unterschiede. Der Grund dafür besteht darin, dass die Grundlage dafür, eine Entscheidung für oder wider ein Verfahren fällen zu können, zunächst einmal Alternativen sind. Mögliche Alternativen können jedoch in erster Linie über die Gemeinsamkeiten der Verfahren beschrieben werden. Der reduktionistische Ansatz von Fayyad et al. [FPSS96b] für die Beschreibung von Data-Mining-Verfahren, dient hier als Grundlage für die allgemeine Beschreibung der Verfahren. Ebenso gilt dieses Modell auch der Identifikation allgemeiner Ansatzpunkte für die Kopplung zwischen automatischen und interaktiven Verfahren.

2.3.1 Data Mining und Monitoring

Die explorative Datenanalyse soll die Formulierung neuer, potentiell relevanter und interessanter Hypothesen ermöglichen. Grundlage dafür ist die Entdeckung von Mustern und Zusammenhängen. Han und Kamber [HK00] definieren Data-Mining entsprechend der obigen Definition für den KDD-Prozess, als die Suche nach nicht-trivialen, neuen und potentiell relevanten Mustern in großen Datensammlungen. Westphal und Blaxton [WB98] betonen, dass die Methoden und Techniken des Data-Mining im Allgemeinen gerade auf jene Probleme angewendet werden, für die Lösungen und Verfahren nur schlecht oder gar nicht a-priori formuliert werden können. Suchverfahren, die durch direkte Programmierung umgesetzt werden können, gehören ebenso wenig zum Repertoire des Data-Mining wie etwa Expertensysteme und Inferenzmethoden, die auf analytischen Artefakten höherer Ordnung operieren. Mannila [Man97] grenzt Data-Mining von der konfirmativen Datenanalyse ab. Automatische, explorative Datenanalyse ist in diesem Sinne deckungsgleich mit Data-Mining.
Data-Mining hat als Forschungsgebiet im Laufe der letzten zwanzig bis dreißig Jahre Einflüsse und Techniken aus verschiedenen Forschungsgebieten absorbiert (siehe z.B. [WB98, HK00]). Zu diesen gehören unter anderem *Machine-Learning* und *Mustererkennung*, *Statistik*, *Datenbanken*, *Algorithmentheorie* und *Optimierung*. Die Zuordnung bestimmter Techniken zu diesen Forschungsgebieten ist nicht einheitlich; das gleiche gilt für die Einordnung des Data-Mining selbst. Eine solche Zuordnung ist jedoch für diese Arbeit auch nicht notwendig. Charakteristisch für die explorative Datenanalyse ist, dass von vornherein nicht bekannt ist, ob sich ein Verfahren oder eine Methode eignet, die relevanten Informationen aus einem Datenbestand zu suchen. Die Kriterien für die Charakterisierung eines Funds müssen wäh-

rend der Analyse erst entwickelt werden. Darin unterscheiden sich Data-Mining und andere Methoden für die explorative Datenanalyse vom *Monitoring*.
Westphal [WB98, Seite 12] unterscheidet die Erstellung von Regeln etwa für die Klassifikation von Datensätzen eines gegebenen Datenbestandes von der Anwendung dieser Regeln auf einen Testfall. Der Monitoring-Prozess entspricht gerade der Anwendung dieser Regeln. Es wird darin kein neues Wissen geschöpft, dessen Bedeutung über den Kontext dieses einen Testfalls hinausgeht. Insbesondere können allein daraus keine allgemein anwendbaren Regeln abgeleitet werden.
Nach einem definierten Verfahren werden die Ausgangsdaten des Testfalls lediglich in eine Form umgewandelt, die es erlaubt, die Entscheidungskriterien eines Nutzers auf die so präsentierten Informationen anzuwenden. In seiner einfachsten Form entspricht Monitoring der Messung (bzw. Erhebung) und Präsentation der Werte, die die Grundlage einer folgenden Entscheidung darstellen. Selbst in diesem einfachen Fall setzt dies jedoch die Sicherheit voraus, dass Messwert und Entscheidungskriterien hinsichtlich der Ziele der Entscheidung relevant sind. Die zugrundeliegenden Regeln könnten zum Beispiel das Ergebnis einer vorangegangen formalen Analyse sein, aber auch auf menschlicher Erfahrung beruhen. Monitoring beginnt im Prinzip an dem Punkt, an dem die Gültigkeit und Relevanz der Regeln nicht mehr bezweifelt wird und diese in der Praxis umgesetzt werden.
Viele im alltäglichen Sprachgebrauch so benannte „Analysen“ sind unter diesen Kriterien betrachtet eigentlich Monitoringverfahren. Beispiele dafür sind etwa:

- Berechnung von Unternehmenskennzahlen
- Dopingtests und medizinische Screenings
- Bonitätsprüfung von Kreditnehmern
- Überwachung von Kreditkartentransaktionen

Monitoring-Verfahren werden mit dem aus einer Analyse gewonnenen Wissen konstruiert. Im engeren Sinne bezieht sich der Begriff auf Wissen, das sich stets auf die Fragestellung eines Anwendungsgebietes bezieht. Allgemeiner betrachtet können Monitoringverfahren auch innerhalb einer Analyse eingesetzt werden, wenn Entscheidungen gefällt werden müssen, die den Analyseprozess selbst betreffen. Es ist denkbar, dass beispielsweise die Qualität eines Klassfikationsverfahrens bei der Anwendung auf neue Testdaten durch die Bestimmung einer Kennzahl überwacht wird. Das Monitoring liefert dabei nicht das Kriterium für den Anwender, sondern das Kriterium für die Wahl eines neuen Klassifikatormodells. Analyse und Monitoring unterscheiden sich daher nicht durch die angewendeten Modelle, sondern dadurch, ob ihre Ergebnisse direkt in eine Entscheidung im Anwendungskontext umgesetzt werden können.

2.3.2 Aufgaben von Data-Mining Verfahren

Die Ziele des Data-Mining lassen sich, abhängig von der zu lösenden Aufgabe auf mehrere mögliche Einzelziele (bzw. Problemklassen) übersetzen. Han und Kamber [HK00] unterscheiden dabei:

- *Klassifikation:* Verfahren in dieser Kategorie „lernen“ eine Funktion $f : X \rightarrow Y$. Dabei beschreiben X und Y Unterräume von Attributen in einem Merkmalsraum $M \supseteq X \times Y$. Sofern nicht explizit angegeben, soll hier und im folgenden angenommen werden, dass die Attribute eines Merkmalsraums einen beliebigen Skalentyp haben können. Ein Trainingsdatensatz ist eine Menge von Tupeln auf diesem Merkmalsraum. Die Beziehung zwischen X und Y wird über die Trainingsdaten modelliert. Ziel der Klassifikation ist die verallgemeinerte Anwendung der Funktion auf einen Testdatensatz etwa für die Vorhersage fehlender Daten. Im Vokabular des Machine-Learning handelt es sich um *überwachte Lernverfahren.*

- *Cluster-Analyse:* Verfahren in dieser Kategorie lernen ebenfalls eine Funktion, die den Trainingsdaten X eine Menge von Clustern Y zuordnen. Die Menge der Cluster und damit die Zuordnung zwischen X und Y ist jedoch vorher nicht bekannt. Die Zuordnung repräsentiert die Ähnlichkeiten innerhalb des Trainingsdatensatzes. Dementsprechend handelt es sich hierbei auch um *unüberwachte Lernverfahren.*

- *Ausreisser-Analyse:* Ziel dieser Verfahren ist eine Identifikation von Daten, die anders charakterisiert werden als der (mehrheitliche) Rest der Daten. Bei der Datenanalyse gilt die Anforderung, dass die Kriterien für die Charakterisierung „abweichender“ und „normaler“ Daten nicht vorgegeben, sondern erst während der Analyse gesucht werden müssen.

- *Trendanalyse & Regression:* Wie bei der Klassifikation wird auch hier wird ein funktionaler Zusammenhang zwischen einer Menge unabhängiger Variablen und einer Menge abhängiger Variablen gesucht. Im Gegensatz zur Klassifikation handelt es sich bei der Trendanalyse bei den Variablen um numerische Werte.

- *Multidimensionale Begriffsbeschreibung:* Die formale Charakterisierung anhand verschiedener Datenattribute. Ist zusätzlich die Extension eines Begriffs bekannt (d.h. die Menge seiner Entitäten) kann dieses Ziel als Klassifikationsproblem formuliert werden für einen bekannten Begriff. Ein ebenso wichtiges Ziel ist die Suche nach *neuen* Begriffen, d.h. die Identifikation von Mustern, die unter Umständen mehrfach in anwendungsrelevanten Zusammenhängen auftauchen sowie deren Beschreibung. Die Konstruktion eines Begriffs durch Charakterisierung seiner Eigenschaften und Benennung erweitert ggf. die Ausdrucksmächtigkeit der Sprache der Anwendungsdomäne.

- *Frequent patterns & Association Mining:* Suche nach Kombinationen von Variablenwerten oder Wertemengen in Datensätzen, die signifikant häufiger sind, als zu erwarten wäre, wenn die Variablen paarweise unabhängig wären. Die Assoziationsanalyse stellt Beziehungen zwischen diesen Kombinationen her.

Dies sind die Aufgaben, die charakteristischer Weise im Data-Mining Schritt des KDD-Prozesses durchgeführt werden. Daraus folgt jedoch nicht, dass der Data-Mining Schritt immer der komplizierteste Schritt des KDD-Prozesses sein muss. In vielen Analysen sind die kritischen Schritte eigentlich diejenigen, mit denen das Data-Mining vorbereitet wird. Zu diesen Aufgaben zählen Han und Kamber *(ebd.)*:

- *Aggregation und Diskretisierung:* Aggregation ist eine Verdichtung der Informationen mehrerer Datenobjekte anhand von vorgegebenen oder bestimmbaren Kategorien. Freiheitsgrade von Aggregationsverfahren sind dabei die Wahl der Kategorien und die Wahl der Verdichtungsfunktion. In manchen Fällen können diese direkt oder indirekt aus der Fragestellung der Analyse abgeleitet werden.
- *Auswahl von Attributen:* Einerseits können Datensätze irrelevante oder redundante Attribute enthalten, andererseits gibt es vergleichsweise wenige Data-Mining Techniken (und *keine* Visualisierungstechnik), die über die Dimensionen eines Datensatzes gut skaliert. Ziel ist also die Auswahl weniger Attribute für die Weiterverarbeitung, die möglichst viele relevante Informationen eines Datensatzes enthalten.
- *Dimensionsreduktion bzw. Merkmalsextraktion:* Dimensionsreduktion kann man als Verallgemeinerung der Attributauswahl betrachten. Das Ergebnis der Dimensionsreduktion sind synthetische Attribute, die als Kombination der natürlichen Attribute konstruiert werden. Die Informationen, die in teilweise abhängigen natürlichen Attributen enthalten sind, kann auf diese Weise effektiver reduziert werden.

Kritisch sind diese Schritte aus drei Gründen. Zunächst dienen alle diese Schritte der Datenreduktion. Der erste Grund ist daher, dass ein möglicherweise bedeutender Anteil an Informationen durch diese Schritte aus dem Prozess ausgeschlossen wird, was natürlich auch das Ergebnis der Analyse beeinflusst.
Der zweite Grund besteht darin, dass man gerade bei der explorativen Analyse nicht aus der Fragestellung erschließen kann, wie diese Schritte durchgeführt werden sollen. Selbst bei einer konfirmativen Analyse ist nicht garantiert, dass eine Hypothese ausreichend präzise formuliert ist, um die Wahl und Steuerung entsprechender Verfahren direkt zu begründen.
Der dritte Grund ist schlicht die Tatsache, dass diese Aufgaben komplex sind. Der Fall, dass kein ausreichendes Vorwissen über die Daten vorliegt, begründet jeweils eine eigene Analyse. Beispielsweise kann die Auswahl von Attributen ein Clustering erfordern, um zu bestimmen, welche Attribute möglicherweise ähnlich sind. Die Aggregation wiederum kann als Ergebnis einer deskriptiven Analyse aufgefasst werden. In diesem Sinne stehen die Verfahren, die das Data-Mining vorbereiten, und das Data-Mining selbst in der Anwendung und auch in ihrer Konstruktion in enger Wechselwirkung. Exemplarisch werden Verfahren für die Dimensionsreduktion ebenfalls vorgestellt (siehe Abschnitt 2.3.5.4).

2.3.3 Allgemeine Charakterisierung von Data-Mining Verfahren

Nicht alle Data-Mining Methoden sind geeignet, um alle diese Aufgaben zu lösen. Vielmehr wurden die zugrundeliegenden Modelle in den meisten Fällen für eine Aufgabe entwickelt, so dass die Aufgabe ihr Repertoire von Verfahren in bestimmten Grenzen determiniert. Ziel des Data-Mining ist das Beschreiben oder das „Lernen“ von Funktionen [FPSS96b]. Wenn ein Verfahren aus allen Trainingsdaten eine allgemeine Beziehung zwischen den Variablen *eines* Datensatzes ableitet und die Funktion diese Beziehung beschreibt, kann man von einem prädiktiven Verfahren sprechen. In einem deskriptiven Verfahren wird stattdessen ein Zusammenhang zwischen *mehreren* Datensätzen und anderen analytischen Artefakten hergestellt. Ziel ist eine Abstraktion und damit eine effizientere Beschreibung der in den Daten verborgenen Strukturen.
Anstelle von Funktionen spricht man allgemeiner von *Modellen*. Jedes Data-Mining Verfahren definiert eine Sprache bzw. eine Modellfamilie, innerhalb der die Modelle beschrieben werden können *(ebd.)*. Betrachtet man die Schemata der möglichen Ein- und Ausgabedaten als gegeben und unveränderbar, sind manche Verfahren mächtig genug, um in diesem Rahmen *jede* denkbare Beziehung zwischen Ein- und Ausgabedaten beschreiben zu können. Mit einer unbeschränkt komplexen Beschreibung wäre man dann in der Lage, die Trainingsdaten akkurat zu repräsentieren. Eine solche Beschreibung ist jedoch nicht hilfreich. Im Data-Mining ist dieses Problem bekannt als *Überanpassung („Overfitting“)* (siehe u.a. Schaffer [Sch93]), wenn die gefundene Funktion zusätzlich zu den eigentlich interessanten Informationen auch das Rauschen in den Daten beschreibt.
Breiman [Bre01b] definiert es als „Ockhams Dilemma“, dass keine formale Sprache in der Lage ist, in jedem möglichen Fall gleichzeitig das genaueste und einfachste Modell zu repräsentieren. Durch die zusätzliche Bedingung, dass ein Ergebnis nur die relevanten Informationen des Datensatzes beschreiben darf, kann die Komplexität der Beschreibung eines Modells beschränkt werden. Dies gilt unabhängig davon, auf welche Weise letztlich bestimmt wird, wie die Informationen vom Rauschen getrennt werden können. Die Wahl einer Modellfamilie wird daher unter der Annahme getroffen, dass sie ein Modell enthält, mit dem die relevanten Informationen des Datensatzes akkurat beschrieben werden können. Die Wahl eines Verfahrens wird zusätzlich unter der Annahme getroffen, dass es die relevanten und die irrelevanten Informationen bestmöglich separieren kann (siehe Abbildung 2.7).

Fayyad et al. [FPSS96a] definieren ein Muster als einen „Ausdruck in einer formalen Sprache oder ein Modell, das eine Teilmenge aller Datensätze beschreibt“. Nach einer anderen Definition von Hand et al. [HSM01] unterscheiden sich Muster von Modellen durch ihren Gültigkeitsbereich: Ein Muster ist ein lokales Merkmal der Daten und umfasst eine Menge an Datensätzen und/oder Attributwerten. Die Existenz eines Musters alleine erlaubt keine Rückschlüsse auf die Verallgemeinerbarkeit der darin enthaltenen Informationen. Ein Modell beschreibt dagegen einen abstrakten, globalen Zusammenhang, der für alle Datensätze gültig ist, d.h. einschließlich aller möglichen Testdaten. Ein Modell beschreibt mithin nicht nur eine endliche Teilmenge aller Datensätze, sondern den sie einschließenden Merkmalsraum.
Entscheidend für diese Arbeit ist jedoch die Tatsache, dass bei automatischen Methoden die prinzipielle formale Beschreibbarkeit eines Musters die Vorraussetzung für das Erkennen eines Musters ist. Wie später gezeigt wird, ist dies bei Visualisierungsmethoden nicht der Fall. Diese Vorraussetzung gilt auch für prädiktive Verfahren, auch wenn bei diesen nicht zwin-

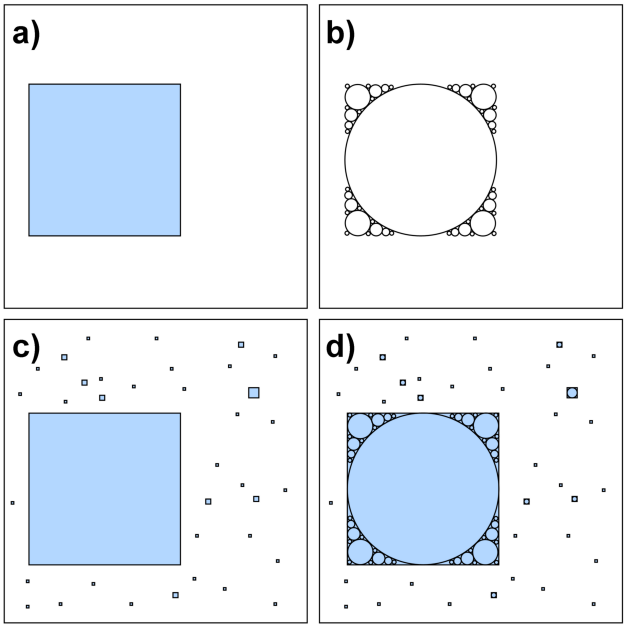


Abbildung 2.7: *Dieses Bild zeigt, wie eng die Auswahl eines analytischen Verfahrens mit der Separation von Mustern und Rauschen innerhalb des Datensatzes zusammenhängt. Bild a) zeigt eine einfache Struktur in der Ebene. Die Menge der Punkte dieses Rechtecks soll ein Muster repräsentieren. Dieses Muster soll durch ein Modell repräsentiert werden. Ein Modell, das Rechtecke beschreiben kann, kann dies leisten, doch diese Auswahl setzt die Kenntnis voraus, dass die Muster in den Daten tatsächlich Rechtecke sind. Bild b) zeigt die Approximation des Musters durch ein analytisches Modell, das nur Mengen beschreiben kann, die Kreise bilden. Auch mit diesem Modell kann eine beliebig hohe Approximationsgüte erreicht werden. Jedoch wird die Relevanz des Modells kompromittiert, sobald der Datensatz auch Rauschen enthält (c) und d)). In diesem Fall erhöht das Verfahren gleichzeitig die Approximationsgüte des Musters und des Rauschens. Durch diese Überanpassung ist schließlich eine Separation zwischen Mustern und Rauschen, und damit relevanter und irrelevanter Information nicht möglich. Allein anhand von quantitativen Fehlermaßen lässt sich die Qualität eines Verfahrens und seines Modells nicht abschätzen. Im Rahmen des hier vorgestellten Konzepts wird die Visualisierung für eine qualitative Analyse des Modells eingesetzt.*

gend erforderlich ist, dass ihre Ergebnisse auch durch den Menschen gelesen werden können. In jedem Fall gilt aber, dass die Wahl eines Verfahrens, seiner Modellfamilie und die Notwendigkeit, die Ausdrucksmächtigkeit zu begrenzen, die Menge von Mustern determiniert, die letztlich gefunden werden können. Dies gilt auch unter der Annahme, dass die Heuristik, mit der die Modelle an die Daten angepasst werden, im Rahmen dieser Begrenzung genauest mögliche Ergebnisse liefert.

Die Unterscheidung zwischen prädiktiven und deskriptiven Modellen ist nicht eindeutig und hängt teilweise davon ab, wofür die Modelle eingesetzt werden. In verschiedenen Arbeiten werden zusätzlich *symbolische* von *subsymbolischen* Verfahren unterschieden (siehe u.a. Chen [Che95], Wnek und Michalski [WM94] und Yao [Yao03]). Diese Charakterisierung „erbte“ Data-Mining vom Forschungsgebiet der künstlichen Intelligenz. Symbolische unterscheiden sich von subsymbolischen Verfahren in erster Linie dadurch, dass die Modellparameter, durch die ein Modell innerhalb seiner Familie definiert wird, eine Bedeutung jenseits des Modells besitzen. Beispiele für symbolische Modelle sind Entscheidungsbäume, Systeme der Aussagenlogik oder Assoziationsregeln. Sie zeichnen sich häufig dadurch aus, dass sie durch den

Menschen direkt gelesen und interpretiert werden können.
In *subsymbolischen* Modellen existiert keine unmittelbare Übersetzung zwischen Modellparametern und Konzepten der natürlichen Sprache. Beispiele hierfür sind künstliche Neuronale Netze, Supportvektor-Maschinen oder die Hauptkomponentenanalyse und andere algebraische Verfahren für die Dimensionsreduktion und Optimierung.
Die Wahl eines symbolischen oder subsymbolischen Modells hat Auswirkungen auf die Möglichkeiten, die Ergebnisse des Data-Mining-Prozesses unmittelbar in Wissen umzuwandeln [Che95]. Wegen der expliziten Beschreibbarkeit der Ergebnisse sind symbolische Modelle dabei deutlich im Vorteil. Wnek et. al [WM94] bemerkt in einem Performance- und Qualitätsvergleich symbolischer und subsymbolischer Techniken, dass die getesteten symbolischen Techniken in der Erkennung einfach strukturierter Muster klar im Vorteil sind, dieser Vorteil aber schwindet, wenn die Muster algebraisch schwer zu beschreiben sind oder Rauschen enthalten. Grundsätzlich lässt sich sagen, dass subsymbolische Modelle in ihrer Ausdrucksmächtigkeit weniger eingeschränkt sind, da das Vokabular ihrer Sprache nicht an die natürliche Sprache gebunden ist. Dies erkauft man sich mit dem Nachteil, dass deren Lesbarkeit mit steigender Komplexität schnell abnimmt.
Breiman [Bre01b] stellt klar, dass mit symbolischen und subsymbolischen Modellen unterschiedliche Ziele verfolgt werden, und dass die Modelle im Bezug auf diese Ziele gewählt werden müssen. Breiman betont den Wert subsymbolischer Modelle mit dem Argument, dass die Genauigkeit eines Modells im Gegensatz zu seiner Interpretierbarkeit das eigentliche Ziel der Analyse sein muss. Diese extreme Forderung wurde *(ebd.)* durch Cox in dem Sinne relativiert, dass eine deskriptive Analyse notwendig ist, um ein Verständnis für die in der Analyse erfassten Zusammenhänge zu erwerben. Als „Black-Box“ Prozess erschließt die prädiktive Analyse keine neuen Zusammenhänge. Auch in dieser Arbeit wird der Standpunkt eingenommen, dass sich prädiktive und deskriptive Zielsetzungen in der Analyse einander ergänzen.

2.3.4 Komponenten von Data-Mining-Verfahren

Nach Fayyad et al. [FPSS96b] erscheint die Menge an verfügbaren Verfahren für die Fragestellungen im Data-Mining oft unüberschaubar groß. Allerdings betonen sie, *„dass der Vielzahl an unterschiedlichen Verfahren, die in der wissenschaftlichen Literatur beworben werden, letztlich vergleichsweise wenige fundamentale Techniken und Modelle zugrundeliegen“*. Um die Data-Mining Verfahren allgemein zu beschreiben, reduzieren sie die Verfahren auf drei Komponenten:

- Gütekriterien für die Qualität des Modells
- Heuristik für die Suche nach Modellparametern
- Modellrepräsentierung

Diese drei Hauptkomponenten sind nicht starr voneinander abhängig. Der Wert dieser Abstraktion erweist sich dadurch, dass man viele in der Literatur beschriebene Verfahren entweder als Variation einer oder mehrerer dieser Komponenten beschreiben kann, oder aber

als (Re-)kombination verschiedener existierenden Techniken für jeweils eine dieser Komponenten [1]. Die Unterscheidung dieser Komponenten erlaubt darüber hinaus auch eine klare Abgrenzung grundsätzlich unterschiedlicher Typen von Parametern, die in einem Verfahren vorkommen.
Modellparameter sind genau die Parameter, die durch das Verfahren gesucht und beschrieben werden sollen. Insbesondere charakterisieren diese ein Modell - das Ergebnis des Verfahrens - innerhalb der Modellfamile. Die Werte der Modellparameter sind vorher nicht bekannt. Davon zu unterscheiden sind die Parameter, die die Gütekriterien und die Steuerung der Heuristik definieren und im Folgenden unter dem Begriff *Verfahrensparameter* zusammengefasst werden sollen. Die Werte der Verfahrensparameter müssen natürlich vorher bekannt sein. Sie sind daher meist auch die Ansatzpunkte für die interaktive Steuerung des Verfahrens. Die Wahl dieser Verfahrensparameter setzt Wissen voraus, dass von vornherein bekannt und gesichert sein muss (etwa als Expertenwissen von Analysten), oder innerhalb der Analyse erst gefunden werden muss - wie beispielsweise in einer Vorberechnung, wie sie in parameterfreien Verfahren durchgeführt wird.
Im Prinzip lässt sich die Auswahl der Modelle, Gütekriterien und Heuristiken als übergeordnete Parametrisierung eines generischen Data-Mining Schrittes auffassen. Die Unterscheidung dieser Parameter widerspiegelt sich in den Methodiken, die von Fayyad et al. und beispielsweise auch den Autoren des CRISP-Modells vorgestellt werden. In einem Data-Mining Schritt innerhalb des KDD-Prozesses werden zwei voneinander abhängige, aber verschiedene Suchprozesse umgesetzt [FPSS96a]. Die übergeordnete Suche entspricht der Bestimmung geeigneter Verfahren und Verfahrensparameter, die nachgeordnete Suche ist die Suche nach optimalen Modellparametern. Ein Ansatzpunkt für das Konzept dieser Arbeit ist der Umstand, dass die Suche nach Modellparametern stets automatisch mit Hilfe der Heuristik durchgeführt wird, die Suche nach den Verfahrensparametern aber (explizit oder implizit) als nicht-automatischer Prozess beschrieben wird.
Nach dem reduktionistischen Modell lassen sich alle Verfahren prinzipiell über die Wahl von Gütekriterien oder Heuristiken steuern (siehe Abbildung 2.8). Die Art der Daten und die Fragestellung der Analyse schränken dabei die einsetzbaren Verfahren meist ein. Häufig ist es jedoch möglich, ein Verfahren oder eine seiner Komponenten durch jeweils kompatible Pendants auszutauschen. Diese Möglichkeit ist letztlich der Grund für die Vielfalt verschiedener Techniken und Varianten.

Die Gütekriterien sind im Prinzip selbst Verfahrensparameter, die abhängig von den zugrundeliegenden Daten die Qualität der Modellparameter beschreiben. Beispielsweise werden Modellparameter - auch in deskriptiven Modellen - häufig danach gemessen, wie gut sie die Referenzdaten in Trainingsdatensätzen approximieren. Gleichzeitig können beliebige Randbedingungen (etwa „Kosten“) in die Gütekriterien eingebaut werden. Die Beschreibung der Gütekriterien erfolgt selbst nach einem Modell - in diesem Fall jedoch eines, das bereits vorhandenes Wissen repräsentieren muss. In der kritischen Analyse stünde aber auch dieses Modell, wie alle anderen Gestaltungsoptionen für ein Verfahren zur Disposition. Ein Beispiel für ein univariates Gütekriterium sind Schwellwertparameter für die Approximationsqualität

[1] Hand et al. [HSM01, Seite 142] beschreiben die gleichen Komponenten, zählen jedoch noch das Datenmanagement zu den Designparametern von Data-Mining Verfahren. Spezielle Datenmanagementsysteme sind notwendig für hochskalierende Verfahren und Heuristiken. Da dies jedoch keinen Aspekt darstellt, der im Konzept dieser Arbeit betrachtet wird, wird dieser Designparameter nicht im Detail beschrieben

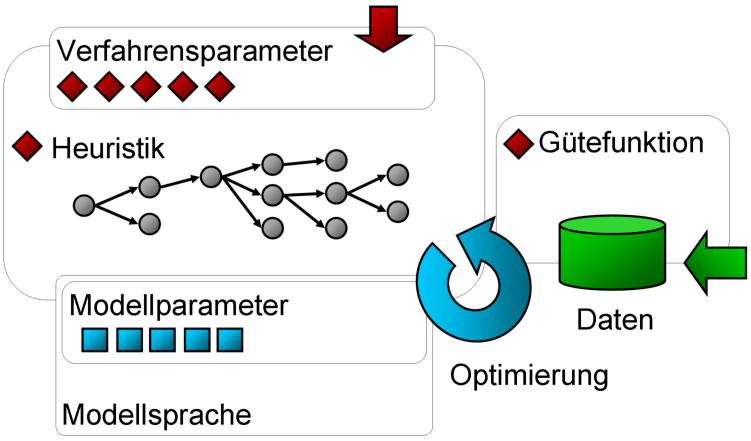


Abbildung 2.8: *Diese Bild zeigt die drei Basiskomponenten eines Data-Mining Verfahrens (nach Fayyad et al. [FPSS96a]): das Modell, die Heuristik und die Gütefunktion. Die Anwendung eines Data-Mining Verfahrens ist insgesamt eine zweistufige Heuristik. Die erste Stufe betrifft die Wahl des Verfahrens und seiner Verfahrensparameter (rot). Prinzipiell kann man auch die Heuristik eines Verfahrens, wie auch die Gütefunktion als Ergebnis dieses ersten Auswahlprozesses betrachten. Dieser läuft in der Regel nicht automatisch ab (vgl. KDD-Modell). Die zweite Stufe ist ein Optimierungsprozess, in dem abhängig von Daten und Gütefunktion die Modellparameter bestimmt werden. Die Modellparameter sind dabei genau jene Komponenten des Modells, die nicht von vornherein bestimmt sind. Die Teile eines Modells, die von stattdessen von vornherein definiert sind, bestimmen, wie die Modellparameter in der gegebenen Sprache interpretiert werden müssen.*

von Entscheidungsbäumen. Dieser Parameter steuert das sogenannte *Pruning* [Sch93], das die Komplexität des Baums begrenzt und Überanpassung vermeiden soll. Ein komplexes Gütekriterium wäre ein Ähnlichkeitsmaß über multivariaten Daten, in dem zum Beispiel sowohl das Modell, das die Ähnlichkeit beschreibt, als auch Gewichtungsfaktoren für die einzelnen Attribute von vornherein festgelegt werden müssen.
Folgende Probleme müssen bei der Wahl aller Verfahrensparameter berücksichtigt werden: Sie haben unter Umständen einen signifikanten Einfluss auf das Ergebnis, der weder kontrolliert noch qualitativ beschrieben werden kann (abhängig von der Komplexität der Verfahren). Dabei besteht gleichzeitig die Gefahr, dass das Ergebnis instabil wird bezüglich der Wahl der Parameter oder dass die Parameter durch den Analysten in Erwartung bestimmter Ergebnisse oder unter bestimmten Annahmen angepasst werden. In diesem Fall, wäre das Ergebnis einer Analyse ebenso ein Artefakt der Parameter wie auch der Daten. Dies ist per se kein Problem, da es zunächst nichts anderes bedeutet, als dass Vorwissen durchaus in Form von Gütekriterien in das Verfahren investiert werden kann und sich auch im Ergebnis manifestiert. Problematisch ist dies allerdings, wenn

- ein Ergebnis ohne Angaben dieser Anpassungen dargestellt wird,
- die Bedeutung der Daten für das Ergebnis von der Bedeutung der Gütekriterien und anderer Verfahrensparameter nicht mehr getrennt werden kann,
- die Bedeutung der Gütekriterien prinzipiell nicht oder nur schwer interpretierbar ist (etwa bei einem subsymbolischen Modell),

- der Anpassungsprozess eines Verfahrens bei der Analyse anderer Daten mit anderen Parametern ausgeführt wird, und die Ergebnisse ohne Berücksichtigung der unterschiedlichen Kriterien verglichen werden.

In jedem der vier Fälle geht Vorwissen, das in die Analyse investiert wird, verloren und wird dabei nicht zuletzt auch dem kritischen Diskurs entzogen. Dabei ist es für die Analyse legitim und oft auch schlicht notwendig, dass bestimmte Konfigurationen einfach ausprobiert werden, ohne dass sie durch Vorwissen gestützt werden. Die Analyse des Parameterraums um dieses Vorwissen zu schöpfen ist eine (Meta-)Analyse eigenen Rechts. Eine Voraussetzung für eine erfolgreiche Metaanalyse ist jedoch schlicht, dass ihre Daten, die ja bei der Analyse erst erzeugt werden, nicht sofort wieder verloren gehen. Unter dieser Voraussetzung besteht mindestens die Chance, dass die Fragestellungen der Metaanalyse mit dem Repertoire an Techniken und Methoden behandelt werden können, *das bereits für die „normale" Analyse zur Verfügung steht.* Dies schließt Visualisierungstechniken ausdrücklich mit ein. Dieser Gedanke wird innerhalb des Konzepts immer wieder aufgegriffen. Daher werden auch jene Teile der Analyse berücksichtigt, die das Data-Mining vorbereiten.

Heuristiken beschreiben Verfahren für die Auswahl und Suche nach Modellparametern, abhängig von den zugrundeliegenden Daten und gesteuert durch die Verfahrensparameter. In vielen Fällen handelt es sich bei den Heuristiken um Optimierungsmethoden, da die Suche nach den Modellparametern über die Gütekriterien im Prinzip als Optimierungsproblem formuliert werden kann. Alle freien Modellparameter bilden einen - im vielen Fällen sehr hochdimensionalen - Suchraum. In der Praxis kann dieser Suchraum nicht erschöpfend durchsucht werden. Wie bei allen Optimierungsverfahren besteht das Kriterium für eine gute Heuristik darin, eine nachweislich optimale oder hinreichend gute Lösung mit möglichst geringem Aufwand zu finden. Von einem rein technischen Standpunkt spielen Heuristiken, im Gegensatz zu den Modellen für das Konzept eine eher nachgeordnete Rolle. Die Verbesserung und Verfeinerung automatischer Heuristiken ist nicht Gegenstand dieser Arbeit; die Verfahren, mit denen der Suchraum durchsucht wird, sollen austauschbar sein, abhängig von den Modellfamilien auf denen sie operieren.
Wichtig für das Konzept ist jedoch die Charakterisierung des Suchraums selbst. Innerhalb des Konzepts sollen bestimmte Varianten einer Steuerung von automatischen Verfahren durch visuell-interaktive Verfahren systematisiert werden. Auf die Heuristiken übertragen bedeutet das die Untersuchung von Methoden, mit denen der Suchraum durch die Information, die der Anwender über die Visualisierung in den Prozess einbringen kann, geeignet eingegrenzt werden kann. Der Suchraum einer Heuristik ist dabei in erster Linie abhängig von dem Modell. Allerdings unterscheiden sich auch automatische Verfahren für die gleiche Modellfamilie häufig darin, welche Einschränkungen sie gegenüber eines potentiell größeren Suchraums machen.
Ein Beispiel dafür sind die unterschiedlichen Verfahrensvarianten für die Konstruktion von Entscheidungsbäumen (zum Beispiel CHAID [PB80], CART [BFOS84]; ID3 [Qui87]; C4.5, C5.0 [Qui96]). Auch in diesen Fällen wird der Suchraum eingegrenzt, in diesem Fall jedoch schon bei der Konstruktion des Verfahrens bzw. der Auswahl einer dieser Varianten für die Analyse. Das Konzept dieser Arbeit unterscheidet sich davon auch nicht darin, *ob* eine Eingrenzung des Suchraums vorgenommen werden soll, sondern wann und auf der Basis welcher Informationen.

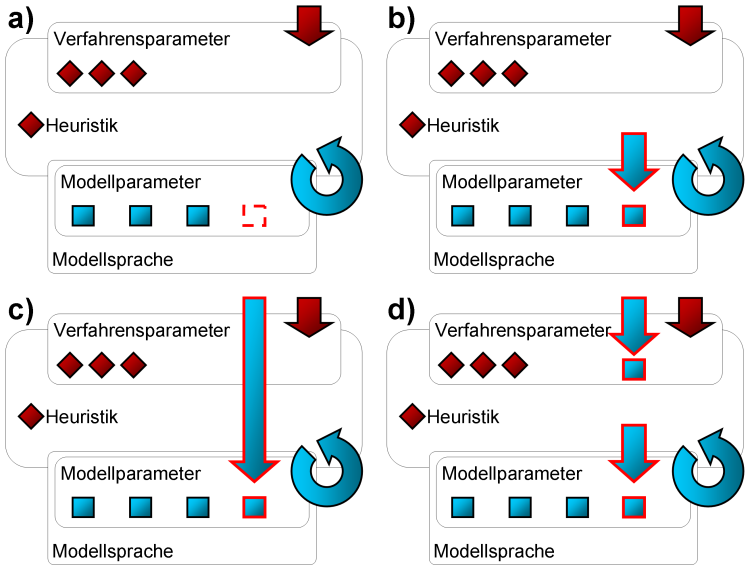


Abbildung 2.9: *Alle Modellparameter können im Prinzip auf unterschiedliche Weise festgelegt werden. Ein Parameter kann durch die Heuristik fixiert werden (a) und ist dann unabhängig von den Daten. Ist der Parameter variabel, kann er normalerweise durch die Heuristik in Abhängigkeit von Daten und Gütefunktionen bestimmt werden (b). Ist der Suchraum sehr groß, besteht jedoch auch die Möglichkeit, bestimmte Modellparameter extern zu definieren (c) oder für die Verfeinerung durch die Heuristik zu initialisieren (d). Besonders durch die letzten beiden Varianten wird es möglich, verschiedene Verfahren - automatische und visuell-interaktive - für die Definition der Modellparameter desselben analytischen Modells zu nutzen.*

Grundsätzlich können die Parameter eines Modells im Suchraum einer Heuristik eine dieser vier Eigenschaften haben (siehe Abbildung 2.9).

- *Parameter ist fixiert*: Der Modellparameter ist allein durch die Wahl der Heuristik vorgegeben, d.h. der Entwickler hat diesen in dem Verfahren vorgegeben. Dies kann eine grundsätzliche Einschränkung der Modellfamilie durch die Wahl der Heuristik bedeuten.

- *Parameter wird intern gesucht*: Der Modellparameter wird nicht vorgegeben und durch die Heuristik bestimmt.

- *Parameter wird extern bestimmt*: Der Modellparameter wird von außen vorgegeben und nicht mehr verändert. Dies ist eine Einschränkung des Suchraums bezüglich der angewandten Heuristik, (d.h. die Heuristik wird durch den Wert des Parameters beeinflusst, sie verändert ihn jedoch nicht).

- *Parameter wird initialisiert*: Eine Kombination zwischen externer und interner Suche. Der Modellparameter wird von außen vorgegeben, aber durch die Heuristik optimiert und dabei ggf. verändert.

Diese vier Fälle beschreiben jeweils eine Kombination der folgenden zwei Eigenschaften, die jeweils zwei Ausprägungen haben: Die erste Eigenschaft beschreibt, ob ein Parameter nur

durch das automatische Verfahren bestimmt wird. Die zweite Eigenschaft beschreibt, ob ein Parameter ein Freiheitsgrad der Optimierung ist und ggf. durch die Heuristik des automatischen Verfahrens verändert wird. Für jeden dieser Fälle lassen sich Beispiele angeben.
Ein Beispiel für einen fixierten Modellparameter gibt es beim CART-Algorithmus für die Konstruktion von Entscheidungsbäumen. Der CART-Algorithmus kann nur Binärbäume erzeugen, wobei das allein nicht die Menge der beschreibbaren Modelle beeinflusst, sondern den Aufwand dieser Beschreibung - CART-Bäume sind im allgemeinen tiefer.
Dazu ist jedoch anzumerken, dass es durchaus eine Frage des Standpunkts ist, ob die Fixierung eines Modellparameters in einer Heuristik als eine Einschränkung einer größeren Modellfamilie verstanden wird, oder ob sie als Modellfamilie eigenen Rechts betrachtet wird. Bei der Untersuchung der Kopplung zwischen visuell-interaktiven und automatischen Techniken ist der erste Standpunkt fruchtbarer. Die Beziehung zwischen verschiedenen Modellfamilien, die durch die Fixierung bzw. Nichtfixierung von Parametern verbunden sind, beschreibt eine *explizite* Designoption für die Modellierung. Auf diese Weise ist es möglich, die Wahl des Analysten zu exponieren und auch zu untersuchen.
Die interne Suche nach (nicht vorgegebenen) Werten für die Modellparameter ist die Aufgabe der Heuristik eines Verfahrens. Beispiele dafür sind etwa die Suche nach Merkmalsvektoren für die Optimierung subsymbolischer analytischer Modelle, aber auch die Suche nach dem besten Attribut für einen Knoten im Entscheidungsbaum.
Ein Beispiel für die externe Bestimmung von Parametern ist der Parameter K des *K-Means*-Clustering Verfahren, der die erwartete Anzahl Cluster angibt. Beim nicht modifizieren K-Means-Clustering wird dieser Parameter nicht verändert; das Ergebnis enthält dementsprechend K Cluster. Eine Vorauswahl der Merkmale oder die Wahl der Distanzfunktion sind ebenfalls Beispiele für externe Parameter. In diese Kategorie fallen alle Parameter, die unabhängig von der Heuristik des Verfahrens bestimmt werden.
Die Unterscheidung zwischen interner und externer Suche wird in dem Fall mehrdeutig, wenn ein Parameter P, der für eine Heuristik $H^1_P : X \to Y$ extern bestimmt werden soll, das Ergebnis eines automatischen Verfahrens mit der Heuristik $H^2 : X \to P$ darstellt. Wenn beispielsweise Entscheidungsbäume auf nicht-diskrete Daten angewendet werden, muss nicht nur für jeden Knoten das beste Klassifikationsmerkmal gefunden werden; um die Qualität eines Merkmals zu testen, muss zusätzlich eine optimale Kategorisierung des kontinuierlichen Wertebereichs bestimmt werden. Die Kategorisierung des Wertebereichs ist ein Clusteringproblem, das mit einer Heuristik H^2 gelöst werden kann, die prinzipiell unabhängig ist von der Heuristik für die Optimierung des Entscheidungsbaums H^1. Mehrdeutig wird die Unterscheidung dadurch, dass $H_{ges} = H^1 \circ H^2 : X \to Y$) insgesamt die Heuristik für die Konstruktion des Entscheidungsbaums darstellt und auch als solche zusammengefasst werden kann. Hinsichtlich der kombinierten Heuristik H_{ges} ist P ein *intern* gesuchter Parameter.
Die Kombinierbarkeit vergleichsweise elementarer Verfahren erzeugt dabei einen komplexen „Algorithmenraum“ (nach Hand et al. [HSM01, Seite 151]). Wie am Beginn dieses Abschnitts erwähnt, ist dies einerseits der Grund für die Vielfalt der überhaupt möglichen Verfahren. Wenn andererseits bei der Konstruktion eines solchen Verfahrens eine solche Kombination fixiert wird, dann wird dadurch eine mögliche Designoption verborgen. Hand et al. betonen *(ebd.)*, dass der Analyst nicht für die Wahl eines kompletten „off-the-shelf“-Verfahrens verantwortlich sein sollte, sondern für dessen Anpassung oder Konstruktion aus entsprechend einfacheren Komponenten. Ein Verfahren H^2 kann in diesem Sinne selbst als Parameter eines

übergeordneten Verfahrens H^1 aufgefaßt werden (siehe Abbildung 2.10).
Die Initialisierung von Parametern kann ebenfalls sowohl über automatische also auch visuell-interaktive Verfahren durchgeführt werden. Der mathematische Hintergrund für die Notwendigkeit einer (nicht-trivialen) Initialisierung ist, dass eine Heuristik auf dem Suchraum nur ein *lokales* Optimum identifizieren kann. Nur eine erschöpfende Suche würde in allen Fällen die Garantie geben, dass ein globales Maximum gefunden wird. Ein Symptom davon ist die Instabilität einer Lösung bezüglich verschiedener Startwerte oder Parameter. Ein Beispiel dafür liefert wieder das *K-Means*-Clustering. So unterscheiden sich mehrere Varianten dieses Verfahrens [Mir05, Seite 86] nur hinsichtlich der Strategien, wie die Startwerte für das eigentliche Clustering berechnet werden.
Wenn eine Heuristik erlaubt, dass Modellparameter extern bestimmt bzw. initialisiert werden können, bietet dies eine Option für die Konstruktion des Verfahrens und des Analyseprozesses. Dabei kann man unterscheiden, ob die Entscheidung durch den Programmierer und Entwickler oder ob die Entscheidung durch den Analysten getroffen wird. Der erste Fall ist insofern kritischer zu behandeln, da die Entscheidungen des Programmierers in der Praxis nicht ad hoc revidiert werden können und in vielen Fällen auch nicht als (potentiell fehlbare) Entscheidungen exponiert werden. Der zweite Fall schließt hingegen sowohl die Möglichkeit ein, dass der Analyst die beste Option kennt, als auch die Möglichkeit, dass er kein Vorwissen besitzt, um die Entscheidungen für die Konstruktion des Verfahrens zu fällen. Durch die Exponierung dieses Freiheitsgrads wird die Datenanalyse durch die Untersuchbarkeit dieses Parameters belastbarer.
Eine Reihe bekannter Analyseverfahren beschreiben eigentlich (Meta-)heuristiken. Eine Beispiel dafür sind etwa *Genetische Algorithmen* (siehe, z.B. Goldberg [Gol89] und Schöneburg et al. [SHF96]). Sie geben im allgemeinen keine Modelle vor, sondern beschreiben eine Klasse von der biologischen Evolution nachempfundenen Heuristiken, die prinzipiell weitgehend unabhängig auf verschiedene analytische Modelle und Gütekriterien übertragbar sind. In der Praxis ist allerdings durchaus so, dass die Qualität der Ergebnisse und die Performanz der Verfahren anhängen von der Abstimmung der Komponenten aufeinander *(ebd.)*.
Weitere Metaheuristiken, sind *Bootstrapping* und *Kreuzvalidierung* (siehe z.B. [Koh95]). Bootstrapping beruht auf einer mehrfachen Anwendung deskriptiver Modelle auf ein Sample aus einer Grundgesamtheit. Ziel ist die Kontrolle der Stabilität des deskriptiven Modells bezüglich der Wahl der Samples, wenn deren Anzahl beschränkt ist. Kreuzvalidierung dient ebenfalls der Vermeidung von Überanpassung und Instabilität bei prädiktiven Verfahren. Test- und Trainingsdaten werden jeweils neu aufgeteilt und daraus werden verschiedene Prädiktoren berechnet. Durch ein Abstimmungsschema werden die Ergebnisse jedes Prädiktors miteinander verbunden.

Die Reduktion des Suchraums für die Heuristiken ist ein zentraler Aspekt, an dem das Konzept ansetzen wird. Das Ziel des Data-Mining ist die Suche nach Mustern in den Daten. Da Muster Mengen sind, ist der Suchraum im Prinzip so groß wie die Potenzmenge der Datenobjekte. Eine solche Suche ist weder praktisch durchführbar, noch würde sie zu einem sinnvollen Ergebnis führen, da sie lediglich eine extensionale Beschreibung des Musters - die Aufzählung aller ihrer Datenobjekte - liefern würde. Eine solche Beschreibung bietet nur einen geringen Gewinn an Information.
Die Modelle für die Beschreibung der Muster stellen eine kompaktere Form der Beschreibung dar, die wesentlich weniger Freiheitsgrade besitzt. Da dies den Suchraum schon erheblich

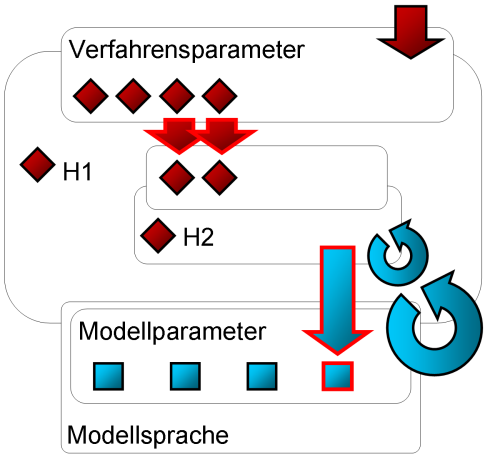


Abbildung 2.10: *Die Komposition von Verfahren ist eine Variante der Kombination verschiedener Techniken. Unterschiedliche Parameter des Modells können dabei durch mehrere geschachtelte Heuristiken optimiert werden. Notwendig wird die Komposition, wenn H^1 eine Metaheuristik wie bspw. ein Genetischer Algorithmus ist. Auch in diesem Fall ist prinzipiell nicht vorgeschrieben, ob die einzelnen Verfahren automatisch oder visuell-interaktiv ablaufen.*

verkleinert, operieren praktisch alle Heuristiken auf den Freiheitsgraden der Modellparameter[2]. Die Suche nach Mustern und die Beschreibung bzw. Modellierung von Mustern sind im Data-Mining daher nicht voneinander zu trennen. Man kann erwarten, dass die meisten automatischen Heuristiken dann eher in ein globales Optimum konvergieren, wenn - bei gleichbleibenden Suchaufwand - der Suchraum kleiner wird, oder wenn die Suche mit „guten“ Startwerten initialisiert wird. Das Konzept dieser Arbeit beruht - aus der Perspektive des Data-Mining - im Kern in der Fragestellung, wie ein Analyst den Suchraum der Heuristik verkleinern bzw. „gute“ Startwerte für die Heuristik identifizieren kann. Dabei gelte die Randbedingung, dass die Heuristiken nicht explizit gesteuert werden sollen. Stattdessen soll die Steuerung aus der Wahrnehmung von Mustern in einer Visualisierung erfolgen.

2.3.5 Modelle und Verfahren

In den folgenden Abschnitten soll ein Überblick über verschiedene Modelle und Verfahren für das Data-Mining gegeben werden. Der Schwerpunkt bei der Beschreibung liegt dabei in erster Linie auf der Charakterisierung der Modelle und ihrer Modellparameter. Wenn Verfahren bzw. Heuristiken für ein Modell näher beschrieben werden, dann dienen sie in erster Linie dazu, Designoptionen für die Gestaltung der Analyse zu identifizieren. Die Techniken selbst werden dann ausführlich beschrieben, wenn sie im Rahmen der technischen Umsetzung des Konzeptes (siehe Kapitel 4) relevant sind. In allen anderen Fällen wird auf entsprechende

[2]Die Suche nach häufigen Mustern ist ein Gegenbeispiel dafür. Mannila [Man97] beschreibt eine generische Heuristik für die Suche innerhalb der Potenzmenge der Datenobjekte. Mannila stellt jedoch an diesem Beispiel auch die Notwendigkeit einer kompakten Beschreibung der Ergebnisse dar, die erst die Grundlage für deren Bewertung und Weiterverarbeitung sein kann.

Überblickstudien verwiesen. Die Aufteilung der Abschnitte folgt dabei der Aufteilung bei Han und Kamber [HK00].

2.3.5.1 Klassifikation und Regression

Algorithmen für die Klassifikation oder Regression optimieren Funktionen $\phi : X \rightarrow Y$. Dabei beschreiben X bzw. Y Unterräume eines Merkmalsraums M von Datensätzen. Ein Merkmalsraum bezeichnet hier allgemein einen Raum von Attributen beliebigen Skalentyps. In den meisten Fällen enthält Y nur ein Attribut dieses Merkmalsraums. Bei der Klassifikation bezeichnet Y eine Menge von Klassen; bei der Regression bezeichnet Y numerische Werte. Die Verfahren erhalten eine Menge von Trainingsdaten $S \subset X \times Y$. Die Zuordnung zwischen abhängigen und unabhängigen Attributen ist durch die Trainingsdaten also von vornherein bekannt. Ziel ist jedoch stets eine Funktion, die nicht nur die Trainingsdaten hinreichend gut beschreibt, sondern eine hinreichend allgemeine Beziehung zwischen allen Daten beschreibt, auf die diese Funktion angewendet werden soll. Der Klassifikator ist demnach ein Modell im Sinne von Hand et al. [HSM01].
Han charakterisiert in [HK00] Verfahren, je nachdem, ob die Heuristik auf die Trainingsdaten angewendet wird, um ein Modell zu erzeugen, dass die Funktion beschreibt (*Eager-Learning*), oder ob die Heuristik auf die Testdaten angewendet wird (*Lazy-Learning*), und dann auf lediglich gespeicherte Trainingsdaten zurückgreift. Zur ersten Kategorie gehören beispielsweise

- Entscheidungsbäume (siehe folgender Abschnitt)
- Bayes-Klassifikatoren (siehe z.B. Heckerman [Hec08])
- Regelinduktion (wie beispielsweise Assoziationsregeln, siehe Agrawal und Srikant [AS94])
- Algebraische Modelle (Lineare Klassifikatoren, Supportvektor-Maschinen, künstliche Neuronale Netze (siehe z.B. [Kra93])

Zur zweiten Kategorie gehören beispielsweise folgende Verfahren

- K-Nearest Neighbor
- Methoden des Fallbasierten Schließens

Bei algebraischen Modellen handelt es sich im Gegensatz zu den anderen genannten Verfahren um subsymbolische Verfahren, in denen jedes Objekt durch einen Merkmalsvektor in einem Vektorraum X repräsentiert ist. X enthält dann im allgemeinen nur numerische Attribute. Die Funktion ϕ wird durch die Verfahren als algebraischer Ausdruck über diesem Vektorraum definiert. Obwohl die entstehenden Modelle bei der Klassifikation und Regression grundsätzlich Prädiktoren sind, lassen sich die Modelle der symbolischen Verfahren häufig auch für die deskriptive Analyse verwenden. Entscheidungsbäume, Bayessche Netze und andere Verfahren beschreiben nicht nur die Zuordnung zwischen den Werten mehrerer Attribute eines Datensatzes, sondern stellen auch die Beziehung zwischen den Attributen selbst dar.

Entscheidungsbäume: Entscheidungsbäume modellieren Entscheidungsregeln, die in einem iterativen Entscheidungsprozess angewendet werden. Sie sind nicht nur aus dem Data-Mining bekannt. Entscheidungsbäume werden auch im Bereich der Wahrscheinlichkeitstheorie, Spieltheorie und in der Entscheidungstheorie als Modell eingesetzt.

Ein Entscheidungsbaum ist eigentlich ein Metamodell. Um einen Entscheidungsbaum zu qualifizieren, genügt es, dass elementare Funktionen - die Entscheidungen - hierarchisch verknüpft werden. Diese elementaren Funktionen beschreiben jeweils eine Partitionierung des Datensatzes derart, dass jedes Datenobjekt stets genau einem Kindknoten zugeordnet werden kann. Entscheidungsbäume sind deshalb Metamodelle, weil jede dieser Partitionierungen selbst ein Klassifikator ist, und es sogar theoretisch möglich wäre, jedes beliebige Klassifikationsmodell in den Baum zu integrieren.
Der Baum definiert bei der Anwendung auf Testdaten einen Pfad zwischen Wurzel und genau einem Blattknoten. Pfad und Blattknoten beschreiben das Profil eines Testdatensatzes und das Prädiktionsergebnis kann im Blattknoten abgelesen werden. Dieses Grundprinzip gilt für alle Entscheidungsbäume unabhängig davon, wie die elementaren Funktionen aufgebaut sind und wie sie jeweils bestimmt werden.
Die übergeordnete Heuristik ist die rekursive Partitionierung des Merkmalsunterraums X für die Konstruktion der Hierarchie. Dieser nachgeordnet können dementsprechend verschiedene Heuristiken und Gütekriterien für die Partitionierung der Teilmenge gewählt werden, die mit jeweils einem Knoten assoziiert ist. In den meisten Fällen sind die elementaren Partitionierungsfunktionen, die den inneren Knoten zugeordnet sind, einfach aufgebaut. Ebenso kommt es fast immer vor, dass diese Funktionen alle aus der gleichen Modellfamilie konstruiert werden. Es gibt keine grundsätzliche Notwendigkeit dafür, aber die Einfachheit hat in der Praxis zwei Vorteile:
Der erste Vorteil betrifft die Optimierung des Baums. Das Grundprinzip für die Konstruktion der Hierarchie selbst steht beim Entscheidungsbaum nicht zur Disposition. Sie gewährleistet allerdings, das die Komplexität eines Prädiktors in viele einfache Entscheidungsschritte zerlegt werden kann, ohne dass dadurch die Ausdrucksmächtigkeit des Modells verringert würde. Je einfacher diese Schritte sind, desto einfacher ist die Suche nach guten Partitionierungsfunktionen. Aus diesem Grund werden sie meist nur auf jeweils ein Attribut des Merkmalsraums X angewendet.
Der zweite Vorteil betrifft die Lesbarkeit der Regel, die der Baum repräsentiert. Die Hierarchie beschreibt nichts anderes als einen Ausdruck, dessen Variablen die Partitionierungsfunktionen darstellen. Durch die hierarchische Zerlegung ist dessen Bedeutung auch Laien gegenüber gut vermittelbar. Die Lesbarkeit des Baums wird daher im Wesentlichen durch die Einfachheit der Partitionierungsfunktionen bestimmt. Im Trade-Off zwischen der Einfachheit der Hierarchie und der Einfachheit der Partitionierungsfunktionen wird daher meist Letztere bevorzugt.
Betrachtet man bekannte Verfahren für die Konstruktion von Entscheidungsbäumen, kann man folgende Unterschiede identifizieren:

- *Anzahl der Partitionen pro Knoten*: Zwei (ID3, CART) vs. variabel (C4.5, CHAID)
- *Anzahl der Attribute pro Knoten*: Univariate Splits vs. Multivariate Splits (CART)
- *Partionierung diskreter und nicht-diskreter Attribute*: (CART, C4.5, CHAID)

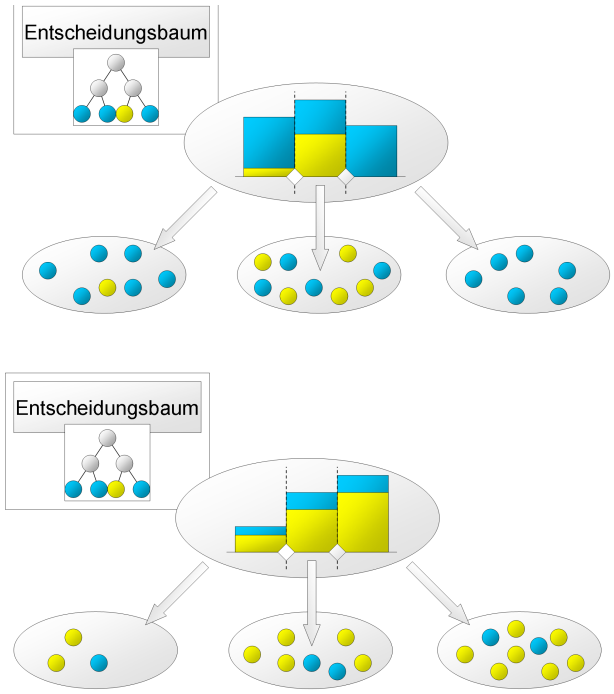


Abbildung 2.11: *Das Gütekriterium definiert welche Partitionierung bzw. welches Attribut für einen Knoten ausgewählt wird. Das Gütekriterium soll beschreiben, wie gut die Partitionierung die Klassen* Y *(hier gelb und blau) auf der Basis eines unabhängigen Merkmals separiert. Das Bild oben zeigt dabei eine vergleichsweise gute, das Bild unten eine schlechte Separation. Die meisten Heuristiken für Entscheidungsbäume basieren darauf, die am besten separierenden Attribute möglich früh im Entscheidungsbaum zu verwenden.*

- *Gütekriterium für Partitionierung*: Chi-Quadrat (CHAID), Entropiemaß (CART, C4.5)
- *Abbruchkriterien*: Pre-Pruning (CHAID), Post-Pruning und andere

Die Abbruchkriterien beziehen sich als einzige unmittelbar auf die Heuristik für die Konstruktion des Baums. Meist wird das Abbruchkriterium aus den Gütekriterien für die Partitionierung abgeleitet. Durch die entsprechenden Pruningstrategien soll eine Überanpassung des Baums an die Trainingsdaten verhindert werden (siehe z.B. Esposito et al. [EMS97] für weitere Details zu Pruningstrategien).
Die anderen Parameter charakterisieren die Partitionierungsmodelle für die inneren Knoten des Baums. Die mit Abstand komplexeste Aufgabe dabei ist die Identifizierung von Unterteilungspunkten („Split-Points“) für die Partitionierung, abhängig von der Anzahl der Attribute und Partitionen. Wie bereits erwähnt, kann diese Aufgabe auf ein Clusteringproblem zurückgeführt werden. Die Charakteristika sind dementsprechend auch gerade solche Parameter, die man für die Definition eines Clusteringverfahrens benötigen würde. Dabei ist zu entscheiden, ob dieses Clusteringproblem einzeln für jeden Knoten gelöst werden soll, oder insgesamt für den gesamten Datensatz. Im ersten Fall muss für den Entscheidungsbaum ein Diskretisierungsverfahren gewählt werden. Im zweiten Fall verlagert man die Aufgabe in die Vorbereitung des Data-Mining.
Zusätzlich zur Wahl der Diskretisierung der Attribute ist die Wahl der Attribute selbst von

Bedeutung. Entscheidungsbäume sind dann besonders gut, wenn man garantieren kann, dass der Merkmalsunterraum X keine abhängigen Attribute enthält. John et al. [JKP94] zeigen die Probleme auf, die entstehen, wenn diese Voraussetzung nicht erfüllt ist.
Selbst wenn garantiert werden kann, dass alle Attribute unabhängig sind, kann ein grundsätzlicher Nachteil von Entscheidungsbäumen auch damit nicht behoben werden. Als Prädiktoren sind sie dann am besten, wenn die Abhängigkeit zwischen den Merkmalsunterräumen X und Y bezüglich der einzelnen Attribute in X separierbar beschrieben werden kann. Beispielsweise sind die Partitionen, die mit univariaten Knoten erzeugt werden können stets achsenparallele Intervalle im Merkmalsraum. Die einfache Beziehung $y = x_1 + x_2$, wobei x_1 und x_2 reellwertige Attribute in X bezeichnen, ist mit einem Entscheidungsbaum mit univariaten Knoten nur mit hohem Aufwand beschreibbar.
Für Entscheidungsbäume wurden verschiedene Gütekriterien etabliert, die meist entweder auf Entropiemaßen (*Informationsgewinn*), oder aber auf statistischen Testverfahren beruhen. Unterschiedliche Kriterien erzeugen dabei nicht notwendigerweise stets den gleichen Baum. Hierfür sei auf die Studie von Ankerst et al. [AEK00] verwiesen, in denen verschiedene Konstruktionsverfahren auf Testdatensätzen untersucht wurden.

2.3.5.2 Clustering

Clustering ist die Suche nach Strukturen in dem Datensatz, die sich aus der Verteilung der Daten innerhalb des Merkmalsraums ergeben. Ziel des Clustering ist eine Reduktion der Komplexität, mit der Nebenbedingung, dass möglichst viel Information über die Verteilung der Daten erhalten bleibt. Ähnliche Objekte werden dabei zu Gruppen, den Clustern zusammengefasst, wobei die Eigenschaften des Clusters die Eigenschaften der einzelnen Objekte möglichst korrekt und spezifisch wieder geben soll.
Nach Berkhin [Ber02] sind Clusteringverfahren aus der Sicht des Machine-Learning Methoden des *Unsupervised Learning*. Im Gegensatz zu den Klassifikationsverfahren ist die Zuordnung $X \times Y$ der Datensätze X zur Wertemenge Y (den Identifikatoren für die Cluster) von vornherein nicht bekannt. Durch das Clusteringverfahren wird diese Zuordnung hergestellt. Diese Zuordnung gilt nicht notwendigerweise auf dem ganzen Merkmalsraum, d.h. sie muss kein Modell im Sinne von Hand sein (siehe Abschnitt 2.3.2).
Grundlage aller Clusteringmethoden und einer der wichtigsten Parameter bei der Konstruktion der Verfahren ist die Ähnlichkeit zwischen zwei beliebigen Datensätzen. Alle Gütekriterien für die Heuristiken beim Clustering sind auf die Ähnlichkeitsmaße angewiesen. Die Wahl einer „guten“ Ähnlichkeitsfunktion hat unmittelbar Einfluss auf die Relevanz der Ergebnisse des Clustering. Die Ähnlichkeitsfunktion ist einerseits abhängig von den Daten und insbesondere den Skalentypen der einzelnen Attribute des Merkmalsraums, andererseits abhängig von der Bedeutung der Ähnlichkeit im Sinne der Problemstellung der Analyse.
Besonders problematisch werden Clusteringverfahren dann, wenn Attribute des Merkmalsraums miteinander verglichen werden müssen, die nicht oder nicht sicher kommensurabel sind. Insbesondere trifft dies auf nominale Attribute zu, für die keine natürliche Distanzfunktion definiert ist. In diesem Fall besteht die Möglichkeit die Distanzen basierend auf der statistischen Werteverteilung (über alle Daten) zu berechnen [HK00], anstelle die Werte direkt zu verwenden.

Auch die Gütefunktion selbst ist beim Clustering nicht notwendigerweise vorgegeben. Diese bestimmt sowohl die Homogenität und Repräsentationsgüte eines Clusters, als auch die Spezifität und Unterschiedlichkeit verschiedener Cluster. Die Distanz zwischen zwei Clustern kann sich beispielsweise auf unterschiedliche Weise aus den Distanzen der Elemente abgeleitet werden.
Clusteringverfahren werden unterschieden nach Berkhin [Ber02], sowie Han und Kamber [HK00] in

- Hierarchische Clusteringverfahren (siehe z.B. Wilkinson und Friendly [WF09])
- Partitionierungsverfahren (z.B. *K-Means*, Mirkin [Mir05])
- Dichtebasierte Verfahren (z.B. *DBScan*, Ester [EKSX96])
- Gitterbasierte Verfahren (z.B. *Clique*, Agrawal et al. [AGGR98])
- Modellbasierte Verfahren (z.B. *Self-Organizing Maps*, Kohonen [Koh97])

Für eine detaillierte Betrachtung dieser Verfahren sei auf die weiterführende Literatur verwiesen.
Die Distanzmaße bestimmten die Topologie auf dem Merkmalsraum, und widerspiegeln in hohem Maße die Annahmen, die über die Daten gemacht werden oder das Vorwissen der Anwender. Sie sind beispielsweise abhängig von den einzelnen Attributen und deren Skalentypen. Für Attribute mit ordinalen und nominalen Skalen gibt es keine natürliche Distanzfunktion; diese muss zunächst konstruiert werden. Selbst bei scheinbar numerischen Attributen wird dabei nicht immer berücksichtigt, dass die Bedeutung der Werte der einzelnen Attribute durch die häufig verwendeten Maße nur schlecht wiedergegeben wird. Beispielsweise ist in einem Anwendungsszenario, bei dem es um die Untersuchung der Häufigkeiten von Kreditkartenbetrug geht, die Abgrenzung von Profilen in denen nie ein Betrug vorliegt wichtiger als die Unterscheidung der relativen Betrugshäufigkeiten. Mit anderen Worten, der Abstand zwischen Null und Eins kann der Bedeutung nach größer sein, als der zwischen Eins und jeder anderen positiven Zahl. Ähnlich konservativ müsste man beispielsweise auch Behandlungserfolge (bzw. Misserfolge) in der Medizin messen. Die resultierende Skala ist in solchen Fällen weder numerisch noch nominal, sondern vielmehr ein Mischtyp.
Viele Clusteringverfahren setzen voraus, dass die Attribute die gleiche Bedeutung haben oder mindestens kommensurabel sind, d.h. die Werte haben einen Bezug zu einer gemeinsamen Skala. In diesem Fall kann eine sinnvolle Gewichtung gefunden werden. Dies gilt dennoch im Allgemeinen nicht für beliebige Attributmengen, insbesondere nicht für nominale oder ordinale Skalen. Selbst für numerische Skalen ist diese Vorraussetzung nicht immer gegeben. Wenn diese Voraussetzung nicht gesichert ist, erfordert die Bestimmung der Abhängigkeiten zwischen den Attributen eine eigene Analyse. Boriah et al. [BCK08] vergleichen verschiedene Ähnlichkeitsmaße für nominale Daten mit dem Ergebnis, dass es keine intrinsisch besten oder schlechtesten Maße gibt. Die Wahl eines Verfahren setzt das Verständnis dafür voraus, wie gut sie die Zusammenhänge und Unterschiede in den Daten charakterisieren.
Die Anwendung einer Distanzfunktion ist ein extremes Beispiel dafür, wie viel Vorwissen über die Daten und die Fragestellung in die Analyse investiert werden muss. Je besser die

Distanzfunktion die Bedeutung der Attribute widerspiegelt, die sie letzten Endes für die Entscheidung des Nutzers hat, desto eher kann man erwarten, dass auch ein automatisches Verfahren „sinnvolle“ Ergebnisse liefert.

2.3.5.3 Ausreißeranalyse

Hodge und Austin [HA04] stellen eine Typologie verschiedener Techniken für die Suche und Modellierung von Ausreißern vor. Ein Ausreißer ist - nach einer der dort vorgestellten Definitionen - „eine Menge von Observationen, die sich inkonsistent zum Rest der Daten verhält“. In der Typologie unterscheiden sie dabei auf allgemeiner Ebene überwachte, unüberwachte und teilweise überwachte Verfahren für die Ausreisseranalyse. Die Unterscheidung überdeckt sich mit der Unterscheidung von Klassifikationsverfahren und Clusteringverfahren. Bei den vorgestellten Verfahren für die Erkennung von Ausreissern handelt es sich entsprechend auch um Varianten dieser Verfahrensgruppen.
Dass die gleichen Verfahren verwendet werden ist nicht verwunderlich, bedenkt man, dass die Identifikation von Ausreissern ein duales Problem zur Identifikation von Mustern darstellt. Mit anderen Worten, um „Inkonsistenz“ beschreiben zu können, muss „Konsistenz“ beschreiben werden, d.h. die „normalen“, prägenden Eigenschaften des Datensatzes müssen formalisiert sein.
Die fundamentale Problematik bei der Ausreisseranalyse ist die Gleiche, wie auch bei der Identifizierung von Mustern bzw. Strukturen: Jedes automatische Verfahren determiniert die Mengen, die als Muster bzw. als Ausreisser identifiziert werden können. Im Rahmen des hier vorgestellten Konzepts wird die Identifikation von Mustern übertragen auf den menschlichen Anwender. Die Ausreisseranalyse wird daher nicht gesondert betrachtet.

2.3.5.4 Dimensionsreduktion und Attributselektion

Ein Beitrag dieser Arbeit sind neue Techniken für die Analyse hochdimensionaler Daten. Eine Vorraussetzung, um diese Daten mit Techniken der Informationsvisualisierung und des Data-Mining bearbeiten zu können, ist die Möglichkeit aus beliebig vielen Dimensionen (bzw. Merkmalen) genau eine minimal kleine Teilmenge zu identifizieren, die für die weitere Analyse noch die nützlichsten Informationen enthält (nach Guyon und Elisseeff [GE03]). Dimensionsreduktion und Merkmalsselektion können als Schritt zur Vorbereitung der Datenanalyse verstanden werden. In Abschnitt 4.2 wird eine Technik für die interaktive Merkmalsselektion vorgestellt, in der das Konzept dieser Arbeit auf diesen vorbereitenden Schritt erweitert wird. Automatische Verfahren aus dieser Klasse werden daher hier detaillierter beschrieben. Dimensionsreduktion wird notwendig, da die Techniken, die für eine Detailanalyse und Mustererkennung eingesetzt werden, im allgemeinen nicht beliebig viele Dimensionen eines Datensatzes gleichzeitig untersuchen können. Dies trifft insbesondere auf Techniken der Informationsvisualisierung zu, aber auch auf Techniken aus dem Data-Mining (siehe beispielsweise John et al. [JKP94]). Eine systematische Untersuchung aller Kombinationen kann der der Größe des Suchraums scheitern.
Ihrem Ziel nach sind Techniken für die Dimensionsreduktion und für die Attributselektion sehr ähnlich: Sie unterscheiden sich darin, dass erstere neue, synthetische Attribute erzeugen

können, letztere aber nicht. Synthetische Attribute haben den Vorteil, dass sie bestimmte Muster in den Daten getreuer repräsentieren können, weil sie eine Kombination mehrere Attribute darstellen. Ihr Nachteil besteht darin, dass sie keine unmittelbar verständliche Bedeutung haben müssen. Zudem sind Techniken für die Dimensionsreduktion nicht in allgemein anwendbar. Praktisch alle Techniken erfordern, dass die Attribute des Datensatzes numerisch sind oder dass mindestens eine Distanzfunktion für nicht-numerische Attribute existiert. Beispiele für diese Methoden sind:

- *Hauptkomponentenanalyse (PCA)* (z.B. Bingham et al. [BGH+06])
- *Projektionsverfolgung* (Friedman und Tukey [FT74])
- *Independent Component Analysis (ICA)* (z.B. Hyvärinen et al. [Hyv99])
- *Neighborhood Preserving Projections (NPP)* (z.B. Pang et al. [PZL+05])
- *Multidimensional Scaling (MDS)* (z.B. Kruskal und Wish [KW78])
- *Self-Organizing Maps (SOM)* (Kohonen [Koh97])

Bei den ersten vier Reduktionsverfahren handelt es sich um lineare Projektionen; das Modell für die Konstruktion synthetischer Attribute ist daher in diesen Techniken gleich. Multidimensional Scaling und Self-Organizing Maps sind dagegen nicht-lineare Projektionen, die zudem auf Distanzfunktionen im Merkmalsraum basieren.
Nichtlineare vektorraumbasierte Projektionsverfahren können als Verallgemeinerung der linearen Verfahren angesehen werden. Während der Bild der linearen Projektion gerade ein linearer Unterraum des ursprünglichen Vektoraums ist, ist das Bild einer nicht-linearen Projektion eine k-Mannigfaltigkeit, wobei k die Dimension des projizierten Raums darstellt. Self-Organizing Maps sind wahrscheinlich das bekannteste Beipiel für diese Art von Techniken. Diese beschreiben eine Klasse künstlicher neuronaler Netze, in denen schrittweise in einem unüberwachten Lernverfahren eine Approximation der Datenpunkte bezüglich dieser Mannigfaltigkeit berechnet wird.
Die im Allgemeinen bessere Approximation erkauft man sich in einer höheren Komplexität der Beschreibung. Während die Topologie der Mannigfaltigkeit in der Form eines Raumgitters vorgegeben wird, sind die zu optimierenden Parameter des Modells die Koordinaten der Gitterpunkte, die als Repräsentanten fungieren. Da die SOM nach der Lernphase auch für solche Daten verwendet werden kann, die nicht mit im Trainingsdatensatz enthalten waren, kann es zu den prädiktiven Verfahren gezählt werden.
Dass Verfahren für die Dimensionsreduktion mit Clusteringverfahren verwandt sind, liegt daran, dass die zugrundeliegende Zielsetzung die gleiche ist. Man kann die Zuordnung von Datenelementen zu Clustern als Konstruktion eines synthetischen Attributes auffassen, mit dem Unterschied, dass die Clusterbezeichner eine nominale Skala beschreiben. Umgekehrt werden Techniken wie die beispielsweise die SOM auch häufig als Clusteringverfahren bezeichnet.
K-Means-Clustering ist ebenfalls ein vektorraumbasiertes Verfahren, dass dem Verfahren

nach große Ähnlichkeiten mit den Kohonen-Karten hat. Den Gitterpunkten der Kohonenkarten entsprechen Clustermittelpunkte (Zentroide) im K-Means Clustering. Der wesentliche Unterschied besteht darin, das die Zentroide nicht durch eine Gittertopologie miteinander verbunden sind und treten daher nur mit den Datenpunkten aber nicht miteinander in Wechselwirkung. Durch iterative Anpassung der Zentroide an die jeweils benachbarten Datenpunkte sollen die Zentroide der in den Daten vermuteten Cluster identifiziert werden. Die gefundenen Parameter des K-Means-Clustering sind die Koordinaten der Clustermittelpunkte. Das Clustering definiert eine Abbildung der Datenpunkte auf den jeweils am nächsten liegenden Clustermittelpunkt. Die Ergebnisse des K-Means-Clustering sind teilweise stark abhängig von der gewählten Anzahl und der Initialisierung der Zentroide. Verschiedene Varianten dieses Verfahrens werden bei Mirkin [Mir05] ausführlich beschrieben.
Kruskal und Wish stellen stellen in [KW78] die Methode des *Multidimensional Scaling (MDS)* vor. Die Eingabedaten sind dabei als Ähnlichkeitsmatrix der Daten gegeben. Ziel des Verfahrens besteht darin, eine Abbildung zu finden, die eine gegebene Stressfunktion minimiert, die das Gütekriterium des Verfahrens darstellt. Die Stressfunktion beschreibt die Unterschiede zwischen „idealen“ Distanzen (d.h. denen, die den Originaldaten zugrundeliegen) und den Distanzen, die tatsächlich durch die Abbildung induziert werden. Innerhalb verschiedener Varianten dieser Verfahrensklasse werden verschiedene Gütefunktionen und Heuristiken eingesetzt.
Die schwächere Bedingung, die an die Attribute des Datensatzes geknüpft wird, wird damit erkauft, dass das Ergebnis des Verfahrens nur eine Abbildung einer diskreten Anzahl von Datensätzen auf Punkte des Vektorraums darstellt. Es handelt sich also nicht um ein Modell im Sinne von Hand et al.. Die Abbildung ist für andere Datensätze nicht definiert und muss bei neuen Daten vollständig neu berechnet werden. Im Prinzip gehört MDS daher zu den Verfahren, die weder zu den deskriptiven noch zu den prädiktiven Verfahren gezählt werden können. Bernataviciene et al. [BDM07] lösen dieses Problem mit einer inkrementellen Variante des MDS, indem die Stressfunktion so modifiziert wird, dass die Abbildung der alten Datensätze fest bleibt, und nur die Position der neuen Datensätze optimiert wird.
Dem gleichen Prinzip folgen die sogenannten Masse-Feder-System für das Layout von Graphen für die Visualisierung (siehe z.B. Herman et al. [HMM00]). Auch beim Layout für Graphen kann den Knoten keine natürliche Position in der Ebene zugewiesen werden, sondern sie muss relativ zu den anderen Knoten bestimmt werden. Dabei muss nicht immer eine ideale Distanz zwischen zwei Knoten angegeben werden. Das Gütekriterium entspricht einer Energiefunktion auf den Kanten, die sich über die Differenz zwischen Ruhelänge und aktueller Länge definiert. Das Layout des Graphen entspricht der Minimierung dieser Auslenkungsenergie auf allen Kanten.
Fodor [Fod02] liefert einen allgemeinen Survey über Verfahren der Dimensionsreduktion. Techniken für die Attributselektion können unter schwächeren Bedingungen eingesetzt werden, und sich daher allgemeiner anwendbar. Insbesondere können diese Techniken auch auf Datensätze mit nominalen oder hybriden Attributen angewendet werden.
Verfahren der Attributselektion enthalten im Wesentlichen eine Heuristik, um eine Teilmenge nützlicher Attribute iterativ zu bestimmen, und eine Qualitätsfunktion, mit der Kandidaten für die nächste Iteration bewertet werden. Kohavi und John [KJ97] unterscheiden für die Qualitätsfunktion sogenannte *Wrapper*- und *Filter*-Methoden.
Wrapper bewerten die Qualität der Attributmenge indirekt über die Qualität der damit

erzeugten Data-Mining Modelle. Beispielsweise kann die Qualität eines konstruierten Prädiktors genutzt werden, um die Wahl der Attribute für die Klassifikation zu verbessern. Die Heuristik für die Attributauswahl und die Heuristik für die Konstruktion des Prädiktormodells werden damit verschränkt. Diese Strategie realisiert eine zielorientiertes Qualitätsmaß, verursacht aber durch diese Verschränkungen einen hohen Aufwand.
Filtermethoden sind stattdessen statistische Schätzmethoden. Bevor ein Data-Mining Algorithmus gestartet wird, werden die Abhängigkeiten zwischen den einzelnen Attributen bestimmt, um daraus die Qualität für die Teilmenge abzuleiten. Dabei können wiederum zwei Fälle unterschieden werden. Im ersten Fall wird die Qualität einzelner Attribute für einen Prädiktor einzeln bestimmt, d.h. unabhängig von allen anderen Attributen (sogenanntes *Ranking*). Hierbei wird in jeder Iteration jeweils das Attribut gewählt, das von den verbliebenen die höchste Relevanz besitzt. Da sich diese zwischen den Iterationen nicht ändert ist die Heuristik sehr einfach und endet bei einer vorgegebenen Anzahl Attribute.
Im zweiten Fall ist die Qualität abhängig von allen Attributen, die in der aktuellen Iteration bereits gewählt wurden - das ist die eigentliche *Feature Subset Selection*. In diesem Fall werden auch Abhängigkeiten zwischen allen Kandidaten berücksichtigt. Die Qualität eines Attributs ändert sich daher in jeder Iteration, was die Heuristik entsprechend komplexer macht. Allerdings können auf diese Weise auch redundante Attribute aus der Auswahl eliminiert werden.
Auch nach der Heuristik können die Verfahren unterschieden werden. Beispielsweise erlauben einige Verfahren das Hinzufügen oder Wegnehmen von Attributen aus der Auswahl oder beides. Sogar Metaheuristiken wie Genetische Algorithmen werden bereits vorgeschlagen (Yang und Honavar [YH98]). Zuletzt können Filtermethoden noch nach der eigentlichen Schätzmethoden unterschieden werden. Eines der gebräuchlichsten Schätzmaße für numerische Daten ist Pearsons Korrelationskoeffizient (siehe, z.B. bei Sirkin [Sir06]), der den Grad der linearen Abhängigkeit zwischen zwei Merkmalen beschreibt. Für eine allgemeine Beschreibung von Abhängigkeiten ist dieses Maß jedoch nicht brauchbar, weil es zum Einen nur eine spezielle Art von Korrelation beschreibt und zum Anderen nicht auf alle Attribute, insbesondere nicht auf Nominale anwendbar ist.
Für nominale Attribute kommen nur solche Qualitätsmaße in Frage, mit denen die Verteilung der Werte geschätzt und verglichen wird. Ein Beispiel für dieses Maß ist die *Transinformation* (engl. *Mutual Information*) zwischen jeweils zwei Attributen:

$$I_{MI}\left(A_1, A_2\right) = \sum_{j=1}^{n}\sum_{i=1}^{m} \underbrace{p(A_{1i}, A_{2j}) log\left(\frac{p(A_{1i}, A_{2j})}{p(A_{1i})p(A_{2j})}\right)}_{:=K(A_{1i}, A_{2j})} \tag{2.1}$$

I_{MI} beschreibt die wechselseitige Abhängigkeit zwischen zwei Attributen A_1 and A_2. Die Mengen A_{1i} und A_{2j} für $i = 1, \ldots, m$ and $j = 1, \ldots, n$ repräsentieren dabei die Partitionen dieser beiden Attributen über denen die Verteilungen verglichen werden. Die relative Mächtigkeit einer Teilmenge der Partitionen im Vergleich zur Gesamtanzahl der Datensätze ist durch p gegeben. Dies ist der Schätzer für die diskrete Dichtefunktion über beiden Attributen.

Brown [Bro09] beschreibt ein System, anhand dessen unterschiedliche Qualitätsmaße unterschiedliche Methoden eingeordnet und verglichen werden können. Browns allgemeine Formel für die Qualität Q eines Kandidatenattributes A_s für einen Prädiktor des Attributes A_t lautet wie folgt:

$$Q(A_s) = \underbrace{I_{MI}(A_s, A_t)}_{\text{Relevanz } Q_{Rel}} - \underbrace{\beta \sum_{i=1}^{s-1} I_{MI}(A_s, A_i)}_{\text{Redundanz } Q_{Red}} + \underbrace{\gamma \sum_{i=1}^{s-1} I_{MI}(A_s, A_i | A_t)}_{\text{Konditional } Q_{Cond}} \tag{2.2}$$

Nach dieser Gleichung unterscheiden sich zahlreiche Methoden in der Gewichtung von Relevanz, Redundanz und Konditional. Diese Unterschiede sind entsprechend beschrieben durch die Gewichte β und γ. Diese einheitliche Beschreibung zeigt zunächst, dass die *Mutual Information* gleichzeitig an mehreren Stellen dieser Formel genutzt werden kann. Zugleich können für eine ganze Klasse von Verfahren auch die Ansatzpunkte für eine Visualisierung und Interaktion einheitlich beschrieben werden. Diese wird in der Realisierung (siehe Abschnitt 4.2.1) auch ausgenutzt.
Zusätzlich zu den entropiebasierten Maßen beschreiben Molina et al. [MBN02] auch eine Reihe andere möglicher Qualitätsmaße. Praktisch alle enthalten eine Summen- oder Integralformel, deren Summanden sich jeweils auf die diskrete oder stetige Werteverteilung beziehen. Die Möglichkeit, diese Summen entsprechend gegebener Partitionen und einzeln zu untersuchen, ist eine Voraussetzung für die allgemeine Anwendbarkeit der später vorgestellten Technik.

2.4 Informationsvisualisierung

Für das Konzept dieser Arbeit stellen die Techniken der Informationsvisualisierung das zu automatischen Verfahren komplementäre Repertoire an Techniken dar. Nach der Definition von Card, Mackinlay und Shneiderman [CMS99] ist Informationsvisualisierung die *Nutzung von computer-unterstützten, interaktiven, visuellen Repräsentierungen abstrakter Daten mit dem Ziel, das Erkenntnisvermögen zu verbessern.* Interaktive Visualisierung ist das Medium für die aktive Auseinandersetzung des Nutzers mit den bearbeiteten Informationen und den Austausch von Informationen zwischen Mensch und Computer.
Der Gesichtssinn ist der Sinn des Menschen mit der mit Abstand größten Bandbreite. Die Funktion des Gesichtssinnes erschöpft sich aber nicht in seiner Eigenschaft als Schnittstelle für die Aufnahme von Informationen; ebenso wichtig noch ist seine Fähigkeit, diese Mengen an Informationen zu verarbeiten, bevor sie kognitiv erfasst werden können (siehe Abschnitt 2.4.4). Bei Erkennung von Strukturen und Mustern ist der Gesichtssinn effizienter und flexibler als jedes automatische Verfahren. Zusätzlich bietet der Gesichtssinn einen strategischen Vorteil: die Erkennung von Mustern kann gelernt werden und kann sich daher anpassen. Generell ist das Gehirn gerade für die Aufgaben, die kognitive Leistung beanspruchen tatsächlich weit weniger gut gerüstet als für solche Aufgaben, die größtenteils nicht kognitiv ablaufen, nicht bewusst wahrgenommen werden und daher scheinbar das Gehirn weniger beanspruchen [Hof01].
Informationsvisualisierung entwickelte sich historisch unter anderem aus folgenden Forschungsgebieten

- *Wissenschaftliche Visualisierung*: Darstellung physikalisch basierter Daten
- *Kartographie*: Darstellung abstrakter und/oder physikalischer Daten im geographischen Raum
- *Datengraphen*: Statische visuelle Darstellung abstrakter Daten

Bertin [Ber83] und Tufte [Tuf83] schufen Systematiken, die die Entwicklung des Forschungsgebietes der Informationsvisualisierung maßgeblich mitbestimmten. Die Wissenschaftliche Visualisierung übernahm seit den achtziger Jahren eine Vorreiterrolle bei der Erschließung der technischen Grundlagen für die interaktive Darstellung am Computer. Zentraler Anspruch in all diesen Forschungsgebieten ist es, Techniken zu entwickeln, mit denen ein möglichst effektiver Austausch relevanter Informationen zwischen Mensch und Maschine unterstützt wird. Ein fundamentaler Unterschied betrifft jedoch die Frage, ob man von vornherein weiß, welche Informationen relevant sind oder ob man dies erst herausfinden will.
MacEachren [Mac95] stellt diesen Unterschied bei der Nutzung geographischer Informationssysteme dar. Grob skizziert sind geographische Informationssysteme im gleichem Sinne eine Weiterentwicklung der Kartographie, wie die Informationsvisualisierung eine Weiterentwicklung der Datengraphen ist. MacEachren ordnet die Ziele bei der Nutzung visueller Repräsentierungen in ein breites Spektrum zwischen *Kommunikation* und *Visualisierung (ebd., 356ff).*

Er charakterisiert die Unterschiede innerhalb dieses Spektrums nach drei Komponenten:

1. Anzahl der beteiligen Nutzer
2. Grad der Interaktion
3. Sicherheit über Korrektheit und/oder Relevanz der dargestellten Informationen

Eine visuelle Repräsentierung für die Kommunikation dient der Vermittlung und dem Austausch von Fakten und Fragestellungen zwischen Menschen. In Rahmen einer Analyse handelt es sich meist um die Präsentation von Ergebnissen. Sie ergänzt damit die übliche textuelle Darstellung und schafft einen gemeinsamen Referenzraum, also eine Ordnungstruktur innerhalb derer die einzelnen Informationen für den Diskurs exponiert werden können. Entlang der gleichen Charakteristika unterscheidet Burkhard [Bur04] den Unterschied zwischen Informationsvisualisierung und Wissensvisualisierung.
Eine visuelle Repräsentierung für die Visualisierung dient dem Austausch zwischen Maschine und Mensch, wobei „Visualisierung“ hier im engeren Sinne nur eine Methode für die Erkundung des Datenraums und die Suche nach neuen Informationen bezeichnet. Auch wenn MacEachren dieses Spektrum für geographische Informationssysteme formuliert, betreffen sie doch visuelle Repräsentierungen im allgemeinen. Im Sinne der Definition von Card et al. soll hier zur Unterscheidung von *Kommunikation* und *Exploration* gesprochen werden, da das in Cards, Mackinlays und Shneidermans Definition genannte Erkenntnisvermögen des Menschen sicherlich durch beide Methoden verbessert werden kann.
Diese Begriffsbildung und Unterscheidung ist von zentraler Bedeutung für das Konzept dieser Arbeit: Beide Ziele - Kommunikation und Exploration - sind für die Datenanalyse gleichermaßen wichtig, wie beispielsweise van Wijk [vW05] ausführt. Entsprechend können Visualisierungstechniken auch für beide Aufgaben eingesetzt werden. Man kann aber nicht erwarten, dass die unterschiedlichen damit verbundenen Anforderungen in einer Technik optimal erfüllt werden. Hier wird die Annahme getroffen, dass Kommunikation und Exploration Aufgaben sind, die sich nicht direkt, sondern nur über ihre Ergebnisse beeinflussen [3]. Wenn es keine Interferenzen gibt, folgt daraus, dass dedizierte Techniken, die genau einer dieser Aufgaben gewidmet und konsequent umgesetzt werden, für diese Aufgabe mindestens so gut geeignet sein sollten, wie nicht spezialisierte Techniken.
Der Schwerpunkt dieser Arbeit liegt nicht auf der Vermittlung bekannter, sondern auf der Erschließung neuer Informationen. Dementsprechend werden die Visualisierungstechniken untersucht, mit denen besonders die Exploration großer und komplexer Datenräume ermöglicht wird. Das vorgestellte Konzept läßt dabei auch eine einseitige Spezialisierung einer bestimmten Technik zu - insbesondere wenn dadurch die Möglichkeiten für die Exploration verbessert werden. Dies wird nur dann möglich, wenn die Funktionen für Kommunikation und Exploration auf unterschiedliche, jeweils spezialisierte Visualisierungstechniken übertragen werden können. Im Rahmen des Konzepts werden bei der Verbindung dieser beiden

[3] Diese Annahme wird durch moderne Szenarien der kollaborativen Analyse eingeschränkt, in denen Kommunikation und Exploration gleichzeitig stattfinden könnten. Allerdings hat die Kommunikation in einem solchen Fall nicht mehr nur die Funktion der Wissensvermittlung und die Beziehung zwischen den beteiligten Akteuren ist symmetrisch: Die Kommunikation, der Austausch von Informationen (anstelle von Wissen), wird zu einem Mittel der Exploration.

Funktionen wiederum automatische Verfahren für die Datenanalyse die entscheidende Rolle spielen.
van Wijk betont *(ebd.)*, dass in der Informationsvisualisierung die explorative Datenanalyse weitaus intensiver und umfangreicher bearbeitet wird als die Präsentation von Daten und Ergebnissen - die folgende Zusammenstellung ist davon keine Ausnahme. Auch die technischen Beiträge dieser Arbeit sind Spezialisierungen für die visuelle Datenexploration. Im Folgenden wird daher das Anforderungsprofil für die entsprechenden Visualisierungen näher erörtert.

2.4.1 Visualisierung für die explorative Datenanalyse

Beim Entwurf von Visualisierungstechniken für die explorative Datenanalyse geht man stets davon aus, dass Nutzer nicht weiß, welche Informationen relevant sind, sondern höchstens weiß, welche Informationen potentiell relevant sind. Durch die visuelle Datenexploration sollen die relevanten Informationen erst gefunden werden. Shneiderman [Shn96] prägte dafür das Mantra *„overview first, zoom and filter, then details on demand“*.
Unter diesem Aspekt betrachtet, gehen die Anforderungen der explorativen Analyse für die Visualisierung weiter als die Anforderung für die Präsentation. Da die Darstellung nicht auf bekannte Informationen zugeschnitten werden kann, muss ein System für die Visualisierung sowohl die Möglichkeiten bereitstellen, *alle* potentiell relevanten Informationen geeignet zu exponieren, als auch die Möglichkeiten dafür, den Suchraum systematisch zu erkunden. Dies hat unter anderem Auswirkungen auf die Art und Weise, wie Wahrnehmung und Interaktion im Analyseprozess eingesetzt werden. Bei der Präsentation soll durch die Art der Darstellung insbesondere die Aufmerksamkeit des Betrachters auf die relevanten Fakten gelenkt werden. Bei der explorativen Analyse würde eine solche Lenkung einen Bias darstellen. Das Mantra kann in den Anspruch übersetzt werden, einen solchen Bias zu vermeiden; allein die Strukturen in den Daten sollten die Aufmerksamkeit des Nutzers leiten.
Dass ein solcher Bias durch eine einzelne Visualisierung vollständig vermieden werden könnte, ist illusorisch. Es gibt keine Visualisierungstechnik, mit der alle möglichen Muster in allen Daten dargestellt werden können. Wie bei automatischen Analyseverfahren, muss man sich beim Design von Visualisierungstechniken der Tatsache bewusst sein, dass dieses Design bestimmt, welche Strukturen und Muster eines Datensatzes sofort, welche nach kognitiver Suche und welche überhaupt nicht gefunden werden können. Wenn man sicherstellen könnte, dass alle Muster in einer Darstellung erkennbar wären, wäre die interaktive Exploration des Datenraums nicht notwendig. Für die explorative, visuelle Datenanalyse ist Interaktion ein integraler Bestandteil des Prozesses, mit dem Anspruch, eine zielgerichtete wechselseitige Beeinflussung des Fokus der Analyse zwischen Mensch und Maschine zu etablieren. Die in dem Mantra kondensierte Vorgehensweise beobachtet Shneiderman in allen Anwendungen.
Shneiderman beschreibt *(ebd.)* eine Taxonomie für fundamentale Datentypen und Aufgaben, die zu den entscheidenden Kriterien für die Entwicklung und Nutzung von Visualisierungstechniken gehören. Das bedeutet insbesondere, dass jede Visualisierung nur für bestimmte Daten und Aufgabenstellungen in der Analyse zugeschnitten werden kann. An Datentypen unterscheidet Shneiderman ein- bis drei-dimensionale Daten, Zeitreihen, Multidimensionale Daten, Hierarchien und Netzwerke.

Ebenso unterscheidet er sieben verschiedene Aufgaben in explorativen Analyse:

- Übersicht
- Zoom
- Filter
- Details
- Historie
- Herstellen von Bezügen
- Extraktion von (Query-)Parametern

Die ersten vier Aufgaben beschreiben dabei am ehesten die Navigation des Nutzers innerhalb des Suchraums. Die Historie beschreibt den Weg, den der Nutzer in diesem Suchraum bereits gemacht hat. Die Exponierung der Historie ist notwendig für die Beurteilung der Systematik des Suchprozesses, mit der letztlich sichergestellt werden muss, dass der Suchraum hinreichend gut exploriert wurde.
Das Herstellen von Bezügen zwischen den Daten beschreibt einerseits die Möglichkeit, die Visualisierung so zu steuern, dass verschiedene Daten innerhalb eines gemeinsamen Kontexts dargestellt werden können. Andererseits beschreibt es die Fähigkeit des Menschen innerhalb dieses Kontexts Muster zu erkennen. Extraktion ist die Transformation von Mustern (d.h. Mengen) in Parametern, die die Muster beschreiben. Shneiderman spricht dabei von Parametern für Suchanfragen (Queries). Da Queries jedoch eine Form von Modellen sind, kann allgemein darunter jede Abstraktion verstanden werden, die eine Menge in der Form von Modellparametern beschreibt. Diese grundlegende Taxomomie von Datentypen und Aufgaben werden in der Folge in weiteren Arbeiten [CMS99, Chi00, AES05, AS05, VPF06] immer wieder aufgegriffen und verfeinert.
Der Unterschied zwischen der Erkennung von Beziehungen oder Mustern und deren Beschreibung ist einer der zentralen Ansatzpunkte für das Konzept dieser Arbeit. Diese Taxonomie lässt die Möglichkeit offen, dass durch eine Visualisierung diese Aufgaben jeweils unterschiedlich gut unterstützt werden. Daraus folgt umgekehrt, dass für eine vollständige Analyse, die alle Aufgaben einschließt, unter Umständen verschiedene Techniken eingesetzt werden müssen.
Praktisch alle Visualisierungsframeworks unterstützen aus diesem Grund das Konzept der *Multiple Linked Views* (siehe z.B. North und Shneiderman [NS00]). Durch die Verbindung der Techniken können verschiedene Visualisierungen für verschiedene Aufgaben eingesetzt werden. Dass innerhalb der Informationsvisualisierung diese Aufgaben auf verschiedene Visualisierungstechniken verteilt werden, ist nicht überraschend. Im Konzept dieser Arbeit wird stattdessen untersucht, wie die Erkennung von Beziehungen durch visuell-interaktive, die Extraktion von Queryparametern (d.h. die Beschreibung von Mustern) jedoch durch automatische Verfahren unterstützt werden kann.

2.4.2 Modelle der Informationsvisualisierung

Eines der einflussreichsten Modelle der Informationsvisualisierung ist der Informationsvisualisierungsprozess von Card, Mackinlay und Shneiderman [CMS99]. Es handelt sich dabei um ein Datenflussmodell, dass die Verarbeitung von Daten in ihrer Rohform bis zum angezeigten Bild darstellt (siehe Abbildung 2.12). Eine Visualisierung wird dabei als eine Serie von - teilweise unabhängigen - Transformationen beschrieben. Der Wert des Modells besteht darin, dass damit die Komplexität des Visualisierungsprozesses in verschiedene Teilkomponenten gegliedert werden kann.
Dem Datenfluss komplementär gegenüber steht die Interaktion des Nutzers. Beide gemeinsam beschreiben einen iterativen Prozess, wobei jede Komponente der Datenverarbeitungspipeline einen Ansatzpunkt für die Steuerung des Prozesses darstellt. Dieses Modell bietet einen Orientierungspunkt aus der technischen Perspektive. Viele Forschungsbeiträge - insbesondere solche, die sich mit technischen Aspekten auseinandersetzen - können innerhalb dieses Modells scharf lokalisiert werden. Der Informationsvisualisierungsprozess ist ein reduktionistisches Modell, das es erlaubt, Verfahren als Kombinationen einzelner Komponenten zu beschreiben und zu entwickeln. Nicht zuletzt dient dieses Modell auch als Grundlage für die Umsetzung von Techniken in der Praxis und das Entwicklung von Softwarearchitekturen für Visualisierungsysteme (wie z.B. bei *Prefuse* [HCL05]). Der erste Schritt der Visualisierungs-

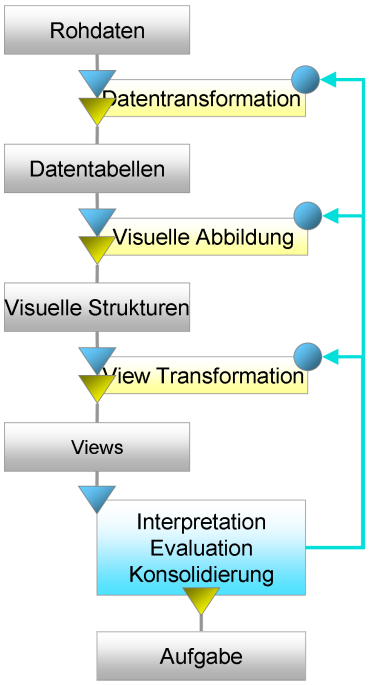


Abbildung 2.12: *Der Prozess der Informationsvisualisierung nach Card, Mackinlay und Shneiderman [CMS99]. Visualisierung und Interaktion sind komplementäre Prozesse der wechselseitigen Beeinflussung von Mensch und Maschine. Die Transformation von Rohdaten in das dargestellte Bild wird in drei - teilweise unabhängige - Komponenten* Datentransformation, Visuelle Abbildung *und* View Transformation *zerlegt. Jede dieser Komponenten kann durch die Interaktion des Nutzers gesteuert werden. Interpretation, Evaluation und Konsolidierung werden durch den Anwender durchgeführt - ähnlich wie auch beim KDD-Prozess (siehe Abbildung 2.6). Sie sind sowohl die Vorraussetzung für die Interaktion als auch für die Wissenschöpfung.*

pipeline ist die Standardisierung der Rohdaten in Datenformate, die mit der Visualisierungstechnik gelesen werden können. Diese Standardisierung wird notwendig, wenn einerseits mehrere verschiedene Datenquellen mit einer Technik angezeigt werden sollen, andererseits auch dann, wenn eine Datenquelle über verschiedene Visualisierungstechniken dargestellt werden soll. Ziel der Datentransformationen ist die Repräsentierung der Daten in einer einheitlichen Struktur und damit eine Entkopplung der Visualisierungstechnik von den Rohdatenformaten. Diese Entkopplung ist die wichtigste Strategie dafür, dass Techniken flexibel auf neue Anwendungsszenarien und Daten angepasst werden können.
Der erste Schritt umfasst dabei nicht nur „triviale“ Transformationen, wie etwa die Umwandlung von einem Datenformat in ein anderes. In vielen Fällen ist es notwendig, unvollständige, ungenaue, fehlerhafte oder sogar falsche Daten zu identifizieren und entsprechend so zu behandeln, dass - je nach Anwendung - die Visualisierung wohldefinierte Daten erhält oder die Visualisierung die Metainformationen selbst geeignet darstellt.
Dieser Schritt der Informationsvisualisierung hat eine direkte Entsprechung im KDD-Prozess (siehe Abschnitt 2.2). Die Datentransformation deckt eine Reihe entsprechender Prozesse in der KDD-Pipeline ab: Datenselektion, Datenaufbereitung und Transformation. Jeder dieser Aufgaben kann in beiden Modellen sinnvoll eingeordnet werden, und jede Technik dafür, selbst wenn sie ursprünglich für eine dieser Technologien entwickelt wurde, ist mit beiden Technologien einsetzbar. Die beiden Modelle für den KDD-Prozess und die das Modell von Card machen keine Aussagen darüber, welches Format die Ergebnisse dieser Prozesse haben. Auf dieser konzeptionellen Ebene sind die Ergebnisse der vorbereitenden Schritte gleich. Die gewählte Form und Beschreibung eines aufbereiteten Datensatzes ist nicht typisch für ein Verfahren aus der Informationsvisualisierung oder des Data-Mining.
In Visualisierungssystemen und Toolkits wie zum Beispiel *Prefuse*, *Polaris* [STH02] oder dem *InfoVis Toolkit* [Fek04] beeinflusst dieser Schritt in besonderem Maße das Design der Softwarearchitektur. Die Standardisierung beschreibt eine Entkopplung der Techniken von den Datenquellen, die den Wiederverwendungswert der Techniken als Softwarekomponenten erheblich erhöht. Die Repräsentierung der standardisierten Daten ist jedoch in verschiedenen System unterschiedlich. Während *Prefuse* Graphen als fundamentale Datenstrukturen nutzt, verwenden *Polaris* und das *InfoVis Toolkit* Tabellen. Dies stellt keine grundsätzliche Einschränkung dar, da diese Repräsentierungen auch ineinander überführt werden können. Von der Wahl dieser fundamentalen Repräsentierung hängt jedoch ab, welche Operationen und Verfahren leicht zu beschreiben sind.
Der zweite Schritt der Visualisierungspipeline ist die *visuelle Abbildung* der transformierten Daten in den „visuellen Raum“. Der visuelle Raum kann beschrieben werden über die graphischen Attribute und ihre Ausprägungen, die Informationen in den Daten codieren können. Diese Transformation ist in dem Sinne der charakteristisch für eine Visualisierung, da die verschiedenen Techniken anhand dieses Schrittes unterschieden werden. Card et al. nennen als zentrale Anforderung [CMS99, Seite 23] sowohl bei der Präsentation als auch bei der explorativen Analyse, dass die visuelle Abbildung die Informationen möglichst getreu im Bild wiedergeben muss. Dass dies keine triviale Anforderung ist, dafür gibt es unter anderem folgende Gründe, die auch in dieser Arbeit eine zentrale Rolle spielen:
Zum Ersten hängt die „Wiedergabetreue“ entscheidend davon ab, wie der Mensch Informationen wahrnimmt und weiterverarbeitet. Eine Visualisierung ist dann „gut“, wenn sie es ermöglicht, die Fähigkeiten des Menschen der Aufgabe entsprechend optimal zu nutzen. Jede

Theorie der Informationsvisualisierung muss bezüglich dieses zweiten Schritts in der Pipeline und der dafür eingesetzten graphischen Attribute unter anderem klären, welche visuellen Attribute und Metaphern durch den Menschen in welcher Weise wahrgenommen werden und welche Informationen daraus rekonstruiert werden.
Ebenso relevant ist, welche Wechselwirkungen bei der Wahrnehmung verschiedener visueller Attribute entstehen. Die Beantwortung dieser Fragen gehört jedoch nicht zu den Zielen dieser Arbeit. Die Grundlagen des Konzepts dieser Arbeit, die sich auf existierende Theorien der Wahrnehmung beziehen, werden im Abschnitt 2.4.4 detaillierter behandelt.
Zum Zweiten besteht die grundsätzliche Beschränkung des Gesichtssinnes auf zwei Raumdimensionen. Alle relevanten oder potentiell relevanten Informationen eines Datensatzes müssen auf diese beiden Dimensionen abgebildet werden. Zwar stehen durch die graphischen Attribute potentiell mehr Freiheitsgrade zur Verfügung, allerdings ist die Anzahl der *unabhängig* voneinander wahrnehmbaren Größen trotzdem sehr klein. Dies ist insbesondere ein Problem in der explorativen Analyse, wo in manchen Fällen eigentlich keine Vorauswahl für die dargestellten Datenmerkmale getroffen werden dürfte. Jede Visualisierung hat eine physikalische obere Grenze, was die Anzahl der darstellbaren Dimensionen und die Komplexität der darstellbaren Abhängigkeiten betrifft. Die Skalierbarkeit bezüglich der Anzahl der Attribute ist zwar auch eine Herausforderung für Data-Mining Verfahren, Visualisierungen sind jedoch im Allgemeinen viel stärker eingeschränkt. Viele Visualisierungsysteme entschärfen dieses Problem, in dem mehrere Visualisierungstechniken miteinander verknüpft werden (*Linking & Brushing*, siehe 2.4.3.2). Die im Kapitel 4 vorgestellten Visualisierungstechniken setzen sich mit dieser Problemstellung besonders auseinander.
Während der zweite Schritt der Pipeline die Transformation in den visuellen Raum beschreiben, beschreibt der dritte Schritt im Gegensatz dazu alle Transformationen *innerhalb* des visuellen Raums, der *View Transformation*. Dabei handelt es sich um Transformationen der Repräsentierung eines visuellen Attributes. Häufig verwendete Transformationen etwa sind Kameraeinstellungen wie Rotation und Zoom (für visuelle Raum-Attribute) und Änderungen der Farbtabelle (für visuelle Farbattribute). Die Wahl überhaupt möglicher Transformationen hängt dabei von der gewählten visuellen Abbildung ab.

Das Ergebnis der letzten Transformation ist das wahrnehmbare Bild. Ab diesem Punkt laufen weitere Prozesse im Gehirn des Menschen ab, wobei das Ergebnis dieser Prozesse eine neue Einsicht oder eine neue Interaktion sein kann. Eine Interaktion dient nach diesem Modell in erster Linie der Veränderung der Transformationsparameter in der Pipeline. Der Informationsvisualisierungsprozess ist ein Modell der Visualisierung, das das Zusammenspiel der *technischen* Komponenten eines Verfahrens beschreibt. Das Modell beschreibt jedoch nicht den Nutzer, bzw. reduziert diesen auf eine Komponente des Systems. Das Modell erklärt daher nicht, warum eine Visualisierung funktioniert, und wie der Nutzer die Informationen aufnimmt und verarbeitet.
Chi et al. [CR98, Chi00] verfeinern die Visualisierungspipeline, indem die Transformationen (in der Arbeit *Operationen* genannt) unterteilt werden in solche, die eine Abbildung von einer Stufe der Pipeline auf die Nächste beschreiben *„Transformation Operators“* und solche, die eine Abbildung innerhalb derselben Stufe der Pipeline durchführen *„Stage Operators“*. Operationen sind dabei Nutzerinteraktionen, bzw. gerade die Verarbeitungsschritte, die die direkt oder mittelbar durch den Eingriff des Nutzers gesteuert werden. Mit ihrer Verfeinerung erlauben sie eine strengere Eingrenzung der Funktion von Operatoren. Derselbe Operator

kann unter Umständen mehrere Funktionen haben, abhängig vom Ansatzpunkt des Operators in der Pipeline. Nach Chi kann die vom Nutzer gewünschte Funktion durch die Angabe des Operators eindeutig definiert werden.

In den meisten Fällen lassen sich verschiedene Ansätze und Techniken der Informationsvisualisierung in ein Spektrum zwischen zwei Polen zuordnen. Ein Ende des Spektrums kann man als „datenzentrierte Ansätze", das andere Ende des Spektrums als „aufgaben- bzw. nutzerzentrierte" Ansätze bezeichnen. Sie unterschieden sich darin, welche Informationen einer Entscheidung für oder wider ein Design zugrundegelegt werden. Amar und Stasko [AS05] stellen diese beiden unterschiedlichen Prinzipien für die Theorie der Informationsvisualisierung gegenüber. In daten-zentrierten Ansätzen sind Visualisierungen in erster Linie eine möglichst getreue Wiedergabe der Daten und ihrer Strukturen - in seiner konsequentesten Auslegung ist dieser Ansatz also unabhängig von Zielen des menschlichen Nutzers.
Amar und Stasko verwerfen *(ebd.)* dieses Prinzip mit der Fragestellung, inwiefern entsprechend entwickelte Systeme geeignet sind, um Entscheidungen, Analysen und Lernen effektiv zu unterstützen. Entscheidungsprozesse, Analysen und Lernprozesse gleichen sich darin, dass Nutzerziele sich wesentlich dynamischer ändern können, als Techniken heute Schritt halten können. Sie skizzieren mehrere „analytische Lücken":

1. Orientierung der Theorie an Daten und Repräsentierungen *(engl. orig.: „representional primacy")*
2. Statische Sichtweise auf analytische Aktivitäten
3. Lücke zwischen Repräsentierung und Analyse

Ihr Hauptkritikpunkt am Stand der Forschung ist der Fokus auf Daten und Repräsentierungen und die Fragestellung, wie der Nutzer mit diesen Repräsentierungen umgeht. Die Folge eines primär an den Daten orientierten Ansatzes ist, dass sich die Möglichkeiten von Visualisierung auch genau darin erschöpfen, die Daten und Strukturen möglichst getreu in der Wahrnehmung des Nutzer wiedergeben zu können.
Das Sichten von Daten ist jedoch nur ein erster elementarer („low-level") Schritt. Eine direkte Unterstützung der diesem Schritt zugrundeliegenden „high-level" Aufgabe kann daraus nicht abgeleitet werden. Die Verwendung existierender Taxonomien für analytische Aktivitäten basiert auf der Annahme, dass die Ziele des Nutzers statisch und explizit formuliert sind, was im Umgang mit Analysewerkzeugen selten der Fall ist. Amar und Stasko *(ebd.)* plädieren hingegen stattdessen dafür, das Design auf der Grundlage der Aufgaben des Nutzers zu entwerfen und damit einen Bezug herzustellen zwischen „high-level"-Aufgaben und „low-level"-Aufgaben. Die Kriterien für die Repräsentierung der Daten werden schließlich daraus abgeleitet.
Ihrer Kritik entsprechend stellen Amar et al. in [AES05] auf der Basis einer Nutzungsstudie eine Liste elementarer Aufgaben zusammen. Sie grenzen ihrer Arbeit insbesondere dadurch von früheren Aufgabenmodellen ab, dass sie nicht die Klassifikation von in der Analyse vorkommenden Datentypen als Ausgangspunkt der Suche sehen. Stattdessen werden die Aufgaben auf die Basiskomponenten in von Nutzern verwendeten Strategien zurückgeführt. Zehn so identifizierte Aufgaben, die in vielfacher Form abgewandelt und kombiniert werden

können, wurden so identifiziert und zur Diskussion gestellt. Sie bemerken dabei auch, dass die gefundenen Basiskomponenten analytischer Strategien sich nicht fundamental unterscheiden von jenen, die mit anderen Methoden identifiziert werden konnten. Valiati et al. [VPF06] verfeinern dieses Modell, in dem sie die analytischen Aufgaben verbinden mit der Wahl dafür jeweils geeigneter Visualisierungs- und Interaktionsmethoden.
Amar und Stasko beschreiben die Beschränkungen existierender Visualisierungssystem in zwei Kategorien: Die „Erklärungslücke“ *(engl. „Rationale Gap“)* und die „Lücke in der Weltsicht“ *(engl. „Worldview Gap“)*. Die Erklärungslücke beschreibt den Unterschied zwischen eine Beziehung in den Daten wahrzunehmen und diese Beziehung erklären und bewerten zu können. Die Lücke in der Weltsicht beschreibt den Unterschied zwischen den Informationen, die dargestellt werden, und den Informationen, die dargestellt werden müssen, um daraus unmittelbar eine Entscheidung ableiten zu können.
Jede der beiden Lücken motiviert jeweils drei Aufgaben auf der höchsten Abstraktionsebene, die durch ein Visualisierungssystem unterstützt werden sollen. Systeme, mit denen die Erklärungslücke überbrückt wird, zeigen nicht nur mögliche Erklärungen, sondern benennen auch ihren Gültigkeitsbereich, zeigen also auch auf, welche Entscheidungen auf der Basis dieser Erklärung nicht getroffen werden können. Als Anforderungen werden genannt:

- Exponierung von Unsicherheit
- Konkretisierung von Beziehungen
- Exponierung von Ursache und Wirkung

Systeme mit denen die Weltsichtlücke überbrückt wird, zeigen nicht nur die Beziehungen in den Daten, sondern benennen auch geeignete Repräsentierungen und deren Grenzen. Dafür nennen Amar et al. folgende Anforderungen:

- Bestimmung von Domänenwissen, das für die Interpretation der Daten und Ergebnisse notwendig wird
- Exponierung mehrdimensionaler Beziehungen
- Tests von Hypothesen

Zusammenfassend fordern Amar und Stasko *(ebd.)*, dass ein System die Ergebnisse einer Analyse jeweils im Kontext seiner eigenen Grenzen darstellt. Das Konzept dieser Arbeit beschreibt kein Visualisierungssystem, sondern soll das Repertoire von Methoden für die Datenanalyse erweitern. Dementsprechend wird hier auch nicht der Anspruch erhoben, dass alle diese Anforderungen erfüllt werden. Das Konzept und die Techniken, die hier vorgestellt werden fokussieren sich in erster Linie auf die letzten beiden Anforderungen: Die Exponierung mehrdimensionaler Beziehungen und den Test von (mehrdimensionalen) Hypothesen.
Die Suche und Exponierung nach mehrdimensionalen Beziehungen ist deshalb eine besondere Herausforderung, weil Visualisierungstechniken, aber auch automatische Verfahren, mit der Dimension der analysierten Daten häufig nicht skalieren. Überdies besteht bei Visualisierungstechniken immer das Risiko, dass allein die Darstellung eines Zusammenhangs eine

Wahrnehmung induziert, die eine Verzerrung oder potentiell ungültige Vereinfachung der Wirklichkeit darstellt. Amar und Stasko lassen offen, ob die Suche, Modellierung und Konfirmation von Hypothesen durch visuell-interaktive oder auch automatische Verfahren umgesetzt wird. Die hier vorgestellte Kopplung zwischen visuell-interaktiven und automatischen Verfahren, kann in diesem Sinne als Methodik verstanden werden, (mindestens) zwei Repräsentierungen - die Visualisierung und ein automatisch erstelltes Modell - gegeneinander abzugleichen.
Amar und Stasko beschreiben mit „Unsicherheit“ die Unsicherheit in den Daten und Aggregationen, die sich über alle Schritte der Analyse bis in die Ergebnisse propagiert. Wie eingangs dieses Kapitels erwähnt, betrifft die hier betrachtete Unsicherheit die Wahl der Methoden, Parameter und Modelle. Auch wenn es sich in dabei Amars Terminologie eher um eine Lücke der Weltsicht handelt, läßt sich diese Anforderung hinsichtlich der Ziele in diesem Modell einordnen.

2.4.3 Interaktion

Insbesondere in der explorativen visuellen Datenanalyse ist die Interaktion eine gegenüber der eigentlichen Visualisierung gleichberechtigte Komponente. Interaktion wird notwendig, wenn der Entwickler eines Verfahrens dem Anwender Freiheitsgrade der Steuerung überlassen muss, die nicht a-priori festgelegt werden können, weil die Bestimmung ihrer Parameter von den Daten oder den Zielen der Analyse abhängt. Gemeinsam beschreiben diese Freiheitsgrade den Suchraum, in dem sich der Anwender bewegt.
Pike et al. [PSCO09] unterscheiden im Begriff „Interaktion“ einerseits die Menge der Kontrollelemente, die dem Anwender zur Verfügung stehen, um die Verfahren zu manipulieren. Andererseits etablieren Visualisierung und Interaktion gemeinsam einen Prozess der wechselseitigen Beeinflussung zwischen Mensch und Maschine, der im abstrakteren Sinne eine Heuristik durch diesen Suchraum darstellt. Pike et al. *(ebd.)* betrachten Interaktion als kognitiven Akt, der durch die Maschine erst ermöglicht wird. Dieser findet keineswegs nur zwischen Mensch und Maschine statt, sondern auch innerhalb des Nutzers.
Munzner [Mun09] schreibt, dass das Gros der Arbeiten in der Informationsvisualisierung, die sich dediziert mit Interaktion auseinandersetzen, allein die technischen Aspekte behandelt. Dagegen gibt es vergleichsweise wenige theoretische Arbeiten. Yi et al. [YKSJ07] geben einen detaillierten Überblick über die bisherigen Taxonomien für die Interaktion und beschreiben selbst eine Taxonomie von Nutzeraufgaben, die durch Interaktion unterstützt oder umgesetzt werden sollen. Diese Taxonomie bezieht sich direkt auf die Taxonomie von Amar und Stasko [AES05] und verfolgt die entsprechende Strategie, um Interaktion auf der Basis der Aufgaben und Ziele des Anwenders zu beschreiben, anstelle der Optionen, die durch die Techniken angeboten werden. Nach einem Review existierender Taxonomien und kommerzieller Softwaresysteme identifizieren Yi et al. dazu einen Satz von "high-level“-Interaktionen.

- Auswählen: *„Selektiere etwas als interessant“*
- Exploration: *„Zeige etwas anderes“*
- Umordnung: *„Zeige die Daten in einer anderen Anordnung“*

- Encodierung: *„Zeige eine andere Repräsentierung [derselben Daten]“*
- Abstraktion: *„Zeige mehr oder weniger Details“*
- Connect: *„Zeige Objektbezüge“*

Damit soll den Entwicklern ein Vokabular an die Hand gegeben werden, mit dem man die gewünschte Aufgabe beschreiben kann, und eine Umsetzung dieser Aufgabe bewerten kann. Allerdings ist zu betonen, dass alle genannten Modelle für die Interaktion deskriptiv sind. Eine allgemeine normative Beziehung zwischen Interaktionen auf verschiedenen Abstraktionsebenen und zwischen Interaktion und Visualisierung wird von keinem der Modelle beansprucht; die Taxonomien sollen vielmehr die Grundlage für die Einordnung und Evaluierung existierender Visualisierungssysteme dienen [YKSJ07].

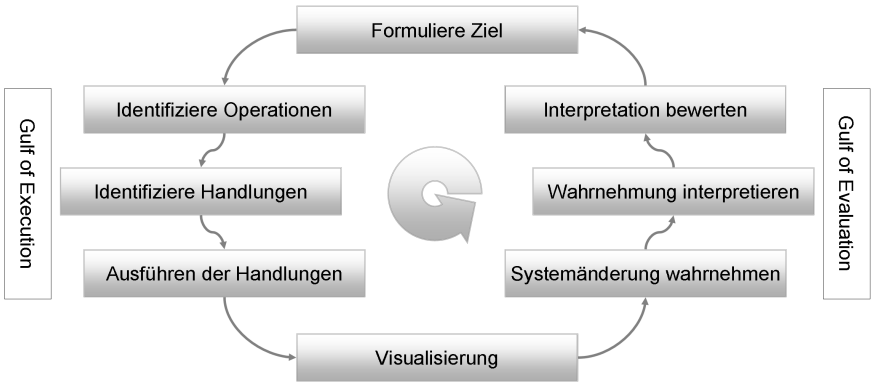


Abbildung 2.13: *Lam [Lam08] überträgt Normans* Aktionsstufenmodell *auf die Informationsvisualisierung als System, mit dem der Nutzer interagiert. Als technisches System verursacht die Interaktion mit der Visualisierung Kosten in jedem der sieben Schritte. In jedem Interaktionszyklus werden dabei alle Schritte durchlaufen. Die Voraussetzung für eine Verbesserung des Interaktionsdesigns ist die Identifizierung und Verbesserung der Stufe(n), in denen die größten Kosten anfallen.*

Lam [Lam08] beschreibt in einer Metastudie, die von spezifischen Techniken für die Interaktion abstrahiert, die „Kosten der Interaktion“. Ihr Modell basiert seinerseits auf dem Aktionsstufenmodell *(engl. orig.: Seven stages of action)* von Norman [Nor02], das allgemein die wechselseitige Beeinflussung zwischen Mensch und Technik beschreibt. Lam verfeinert und formuliert diese sieben Stufen für die Interaktion und Visualisierung um (siehe Abbildung 2.13). Sie charakterisiert existierende Visualisierungssysteme nach ihren jeweiligen Problemstellungen, die entsprechend hohe Kosten verursachen und diskutiert die Strategien, um diese Kosten zu senken. Es unterteilt die Interaktion in die beiden Richtungen der wechselseitigen Beeinflussung. In der ersten Richtung werden aus Handlungszielen des Nutzers eine ausgeführte Handlung für die Steuerung der Maschine, in der zweiten Richtung wird die Wirkung dieser Handlung erkannt, interpretiert und bewertet um ggf. neue Ziele formulieren zu können.

Zusammengefasst ergeben sich zwischen allen sieben Stufen Problemstellungen, die potentiell hohe Kosten verursachen:

1. Entscheidungskosten für die Auswahl einer Teilmenge der Daten oder einer möglichen Option der Nutzerschnittstelle
2. Kosten für die Umsetzung der Entscheidung in eine Sequenz von Aktionen
3. Kosten für die Durchführung der Aktionen
4. Kosten für die Wahrnehmung
5. Kosten für die Interpretation der Handlungswirkung
6. Kosten für die Evaluation

Lam beschreibt in ihrem Modell die Kosten, die durch die Möglichkeit der Interaktion erst geschaffen werden. Gemeint sind dabei eigentlich die Kosten, die durch die Interaktionsmetaphern auf dem Bildschirm (etwa Mauszeiger oder Tooltips) hervorgerufen werden. Die Kosten für die Interpretation der Handlungswirkung entsprechen daher nicht exakt den Kosten für die Interpretation des Bilds selbst (z.B. der wahrgenommenen Muster), sondern denen für die Interpretation der Bild*änderung* nach einer Interaktion. Im Prinzip lässt sich dieses Modell jedoch auch auf die Wahrnehmung und Interpretation von Mustern übertragen, da diese mindestens bei der ersten Interpretation eines Bildes anfallen. Die Erstinterpretation eines (statischen) Bilds wird hier dementsprechend als Spezialfall dieses Modells aufgefasst.
Die Unterscheidung zwischen der Wahrnehmung und Interpretation in Lams und Normans Modellen ist ein zentraler Ansatzpunkt für das Konzept dieser Arbeit. Für die Ableitung potentiell relevanter Informationen aus Mustern sind beide Teilprozesse relevant. Jedoch erlaubt es diese Unterscheidung für jeden der beiden Teilprozesse eine eigene Lösung zu suchen, deren Kosten jeweils unabhängig voneinander minimiert werden können. Eine entsprechende Strategie wird im Konzept mit der Kopplung automatischer und visuell-interaktiver Verfahren beschrieben.
An dieser Stelle muss auch betont werden, dass das in dieser Arbeit verfolgte Ziel nicht in der Verfeinerung der Ausführung der Interaktion besteht, wie es Lam *(ebd.)* einordnet und andere Autoren bereits vielfach umgesetzt haben:
In verschiedenen Arbeiten werden Techniken beschrieben, in denen die Komplexität der Steuerung reduziert wird. Schneidewind [Sch07], wie auch Hao et al. [HDK+07] beschreiben Verfahren für die automatischen Steuerung von Visualisierungstechniken, die auf der Analyse der Nutzerinteraktion mit Data-Mining Verfahren basiert. Bezüge zu anderen Teilen eines Datensatzes werden automatisch identifiziert und automatisch dargestellt.
Jankun-Kelly beschreibt in [JK03] Spreadsheed-basierte Interaktionsschnittstellen, in denen potentiell geeignete Konfigurationen vorgeschlagen werden. In dieser Arbeit wird der Suchraum, innerhalb dem sich die Nutzerinteraktion bewegt, als so genannten *P-Sets* (=Parametermengen) formalisiert. Diese P-Sets beschreiben wiederum die Modellparameter, die als Ansatzpunkt für automatische Heuristiken interpretiert werden können. Dieser Ansatz wird von Jankun-Kelly et al. [JKMG07] in ein Softwareframework für die Interaktionsanalyse erweitert.

Die Vereinfachung der Steuerung orientiert sich in diesen Fällen nicht nach den Freiheitsgraden des Systems, sondern nach den Aufgaben der Anwender. Auch wenn die Interaktionsmodelle von Norman und Lam eine zentrale Stelle im Konzept dieser Arbeit einnehmen, liegt das Ziel dieser Arbeit nicht in der Verbesserung bzw. Vereinfachung der Ausführung einer Interaktion, wie zum Beispiel für die Steuerung der Visualisierung bei die Suche nach Beziehungen im Datensatz.
Das Ziel dieser Arbeit liegt stattdessen in der Senkung der Kosten für die Interpretation des wahrgenommenen Bildes. Für die Interpretation komplexer Muster werden automatische Verfahren genutzt, die diese Muster in ein formales Modell übertragen. Die automatischen Verfahren haben keinen direkten Einfluss auf die Navigation des Betrachters im Suchraum. Die Visualisierung bleibt unverändert und bietet einen Referenzrahmen für den Abgleich zwischen wahrgenommenen Muster, und den Mustern im Datensatz, die durch die Modelle beschrieben werden.
Kosara et al. [KHG03] stellen eine Taxonomie von Interaktionstechniken zusammen. Sie beschreiben im einzelnen folgende Methoden:

- Fokus und Kontext
- Übersicht und Detail
- Filtering und Magic Lenses
- Linking und Brushing

Weitere Interaktionstechniken und Konzepte sind *Interaktives Filtern* [Kei02], *Dynamic Queries* [AWS92] und *Direkte Selektion* (siehe z.B. Derthick et al. [DKR07] oder Shneiderman und Plaisant [SP04]). Das Modell von Card et al. (siehe Abschnitt 2.4.2) erlaubt es, eine Interaktionstechnik danach zu kategorisieren, je nachdem, welche Schritte des Visualisierungsprozess durch die Interaktion direkt beeinflusst werden. Man kann demnach Techniken für die Steuerung der View-Transformation unterscheiden (Panning und Zooming, Focus und Kontext, Overview und Detail), Techniken für die Steuerung der Visuellen Abbildung (Magic Lenses) und Techniken für die Beeinflussung der Datentransformation (Linking & Brushing, Animation).
Diese Kategorisierungen sind jedoch nicht immer kongruent. Keim [Kei02] beschreibt interaktives Filtern als Technik, die sowohl die Datentransformation als auch (potentiell) die Visuelle Abbildung beeinflusst. Heer und Robertson wenden in [HR07] die Animation auf die Visuelle Abbildung an. Relevant für das Konzept dieser Arbeit sind, unabhängig von sonstigen Kategorisierungen, die Techniken für die Manipulation der Datentransformation. Dies gilt insbesondere für die Direkte Selektion und die Dynamic Queries, weswegen sich die Darstellung im Detail auch auf diese beschränken soll. Bei den anderen Interaktionstechniken sei hier auf die angegebene Literatur verwiesen.

2.4.3.1 Direkte Selektion

Direkte Selektion ist genau genommen keine Interaktionstechnik, sondern ein Interaktionskonzept. Es gibt nicht vor, welche Funktion die durch die Interaktion gegebenen Daten

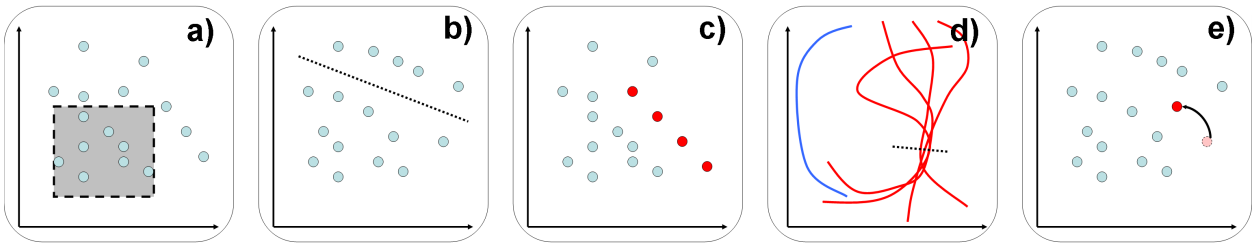


Abbildung 2.14: *Fünf Varianten für die Verwendung direkter Selektion in der Visualisierung. (a) beschreibt die direkte Selektion eines Bildbereichs (siehe z.B. Derthick et al. [DKR07]). Variante (b) kann interpretiert werden als Separation von zwei Mengen von Datenobjekten, oder auch als Separation von Bildbereichen. Um eine Selektion handelt es sich dabei in dem Sinne, dass ein Modell direkt gewählt, dass die Separation beschreibt (in diesem Fall eine trennende Hyperebene). Die Varianten (c) und (d) zeigen die direkte Selektion von Datenobjekten. (e) beschreibt eine „Drag & Drop" Metapher für die direkte Manipulation. Die verschiedenen Varianten können in verschiedenen Visualisierungstechniken eingesetzt werden.*

steuern sollen, sondern auf welche Weise sie mit einem Zeigegerät erzeugt werden. Direkte Selektion findet grob umrissen immer in der Visualisierung statt. Präziser beschreibt „direkt" einen Bezug zwischen dem Ort (auf dem Darstellungsgerät) an dem die Interaktion stattfindet, und dem Ort, an dem der Nutzer wahrnimmt, was durch die Interaktion letztlich ausgelöst wird.
Dieser Bezug ist dann direkt, wenn der Nutzer eine Korrespondenz zwischen diesen beiden Orten herstellen kann. Dies gilt sicher dann, wenn der Ort von Wahrnehmung und Handlung identisch ist, dies ist aber nicht immer notwendig. Dass beispielsweise in einem Koordinatensystem Punkte gleicher Höhe mit der entsprechenden Höhe auf der Ordinate korrespondieren, wird erlernt und internalisiert. Der Anwender leitet die Korrespondenz aus der Konvention über die Metaphorik des Koordinatensystems implizit ab.
Der Unterschied zwischen „direkter" und „indirekter" Selektion ist in diesem Sinne nicht ganz eindeutig. Die Unterscheidung hängt auch davon ab, ob der Anwender dieses Korrespondenzproblem selbst lösen muss, ob dieser Bezug durch Konvention oder Erfahrung bereits hergestellt ist, oder ob beispielsweise visuelle Hinweise die Lösung dieses Problems erleichtern. „Indirekt" ist in praktisch allen Fällen eine Anwendung des Zeigegeräts außerhalb der Visualisierung, die sich jedoch auf die Visualisierung bezieht - im allgemeinen durch die Elemente der graphischen Benutzeroberfläche. Das Forschungsgebiet des User-Interface Designs beschäftigt sich dementsprechend mit der Vereinfachung indirekter Selektion.
Aus dieser Abgrenzung lassen sich Vor- und Nachteile der direkten Selektion ableiten: Die Vorteile bestehen darin, dass Wahrnehmung und Handlung (im Sinne von Lams Modell) unmittelbar aneinander gekoppelt werden können. Um die Gesamtkosten für die Interpretation einer Visualisierung bewerten zu können, ist es notwendig, eine Unterscheidung zwischen erlernten und nicht erlernten Korrespondenzen zu treffen. Im Rahmen des Konzepts dieser Arbeit ist dies deshalb relevant, weil diese Interaktion lediglich auf der *Erkennung* von Mustern basieren muss, jedoch keine „höheren" kognitiven Funktionen beanspruchen darf. Im folgenden Kapitel wird im Detail beschrieben, wie die Trennung von Mustererkennung und Musterbeschreibung auf der direkten Selektion aufbaut. Die direkte Selektion stellt dabei den Kanal dar, über den Informationen über die wahrgenommenen Muster vom Menschen zur Maschine geleitet werden.

Jede Information, die durch direkte Selektion definiert werden kann, ist jedoch direkt von der Visualisierung abhängig. Technisch gründet direkte Selektion auf der Bestimmung des Urbilds der selektierten Bildpunkte. Das Urbild der visuellen Abbildung bezeichnet präziser nach Cards Modell das Urbild der Hintereinanderausführung von visueller Abbildung und View-Transformation.
Derthick et al. [DKR07] unterscheiden zwei Arten von Urbildern, die durch Interaktion definiert werden können. [4] *Extensionale* Urbilder, sind die Ergebnisse der direkten Selektion von Datenobjekten. Sie repräsentieren daher stets eine Teilmenge der dargestellten Datenobjekte, und zwar in der Form einer Aufzählung ihrer Namen (bzw. Identifikatoren). *Intensionale* Urbilder sind die Ergebnisse der anderen Interaktionsformen. Sie repräsentieren eine Teilmenge des Datenraums, d.h. eine Menge von Tupeln der Attribute. [5] In bestimmten Visualisierungen ist die Bestimmung der Urbildmenge der visuellen Abbildung nicht möglich oder nicht effizient. Dies gilt zum Beispiel für nicht-lineare visuelle Abbildungen wie bestimmte Algorithmen für das Layout von Graphen oder die Darstellung von Ergebnissen des *Multidimensional Scaling* (siehe Abschnitt 2.3.5.4). In diesem Fall wäre nur die Wahl von Datenobjekten einsetzbar. Tweedie et al. stellen in [TSDS96] Visualisierungen vor, in denen keine Datenobjekte vorkommen. In diesen Visualisierungen ist infolgedessen nur die Selektion von Bildbereichen einsetzbar.
Die Visualisierung bestimmt daher, welche Informationen durch direkte Selektion aus der Interaktion des Nutzers für die Maschine direkt abgeleitet werden können. Das Konzept dieser Arbeit umfasst daher nicht nur die Fragestellung, welche Rolle die direkte Selektion bei der Arbeitsteilung zwischen Mensch und Maschine für die Identifizierung von Mustern spielt. Es behandelt auch die Frage, ob für die Teilaufgaben der Analyse (im Sinne der Taxonomie von Han und Kamber für Data-Mining Aufgaben 2.3.2) eine Visualisierung bestimmt werden kann, die diese Arbeitsteilung überhaupt erst zulässt.

2.4.3.2 Linking & Brushing

Der Begriff „Linking & Brushing“ beschreibt eine Verbindung zwischen zwei oder mehr Ansichten auf die gleichen Daten. Sie baut auf der direkten Selektion auf. Eine Selektion oder Hervorhebung in einer Ansicht verändert die Darstellung in den anderen Visualisierungen. North und Shneiderman stellen in [NS98] eine Taxonomie verschiedener Methoden für die Koordination zwischen zwei verschiedenen Visualisierungen zusammen. Sie unterscheiden dabei, ob die Interaktion durch die Selektion von Datenobjekten oder durch die Änderung der Ansicht ausgeführt wird, und ebenso, ob die Interaktion die Selektion von Datenobjekten oder die Änderung einer anderen Ansicht bewirkt.
Linking & Brushing ist eine Interaktionsmethode, die die ganze Visualisierungspipeline überspannt, denn sie setzt erstens voraus, dass die gleichen Daten mehrere Visualisierungspipelines durchlaufen, und zweitens, dass in jeder dieser Pipelines Umkehrabbildungen für die jeweiligen Transformationen definiert sind. Es muss möglich sein, zurückzuverfolgen, auf welche Daten oder Wertebereiche sich ein dargestelltes Objekt oder ein Bildschirmbereich bezieht.

[4] Derthick et al. beschreiben eigentlich Suchanfragen; diese korrespondieren jedoch direkt mit den Urbildern der visuellen Abbildung, weswegen hier die gleiche Kategorisierung verwendet werden kann.

[5] Keim et al. [Kei02] unterscheiden entsprechend die Selektion einer Teilmenge der Datensätze (als *„Browsing“*) und der Spezifikation ihrer Eigenschaften (als *„Querying“*).

Innerhalb eines Visualisierungssystems gibt es keine dedizierte Komponente für Linking & Brushing. Vielmehr muss die Funktionalität bereits innerhalb der einzelnen Visualisierungskomponente und dem Datenmanagementsystem angelegt sein. Die Visualisierungstechniken müssen die Funktion zur Verfügung stellen, die einzelne Pixel den Datenobjekten zuordnet, die sie belegen. Das Datenmanagementsystem muß einen einheitlichen Namensraum (bzw. Adressraum) für die Beschreibung der Datenobjekte in allen Visualisierungen zur Verfügung stellen.

Linking & Brushing erhöht den Nutzen einzelner Visualisierungstechniken durch die Kopplung beträchtlich. Im Prinzip handelt es sich dabei um eine Umgehung der natürlichen Beschränkung einzelner Visualisierungen auf wenige Dimensionen. Die gleichen Daten können unter verschiedenen Aspekten betrachtet werden, wobei durch diese Technik ein automatischer Bezug zwischen diesen Techniken hergestellt wird. Ein Beispiel für ein System, das diese Technik implementiert, beschreiben North und Shneiderman [NS00]. *Linking & Brushing* ist weitgehend unabhängig von der oder den eingesetzten Visualisierungstechniken, weil die Verbindung der Techniken über das gemeinsame Datenmodell erfolgt. North formalisiert dabei das Problem, dass das System eine konsistente Datenreferenz bereitstellen muss (siehe Abbildung 2.15).

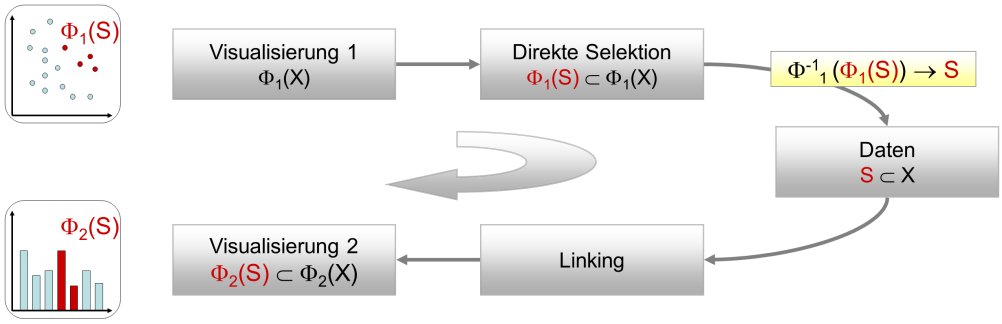


Abbildung 2.15: *Voraussetzung für* Linking & Brushing *ist ein gemeinsames Datenmodell für die verbundenen Visualisierungstechniken* Φ_1 *und* Φ_2*, und die Definition der Umkehrabbildung für jede der Visualisierungen. Da durch* Brushing *eine Teilmenge* S *nicht direkt definiert wird, sondern stets nur ein Bild* $\Phi_1(S)$*, muss dieses Bild über die Umkehrabbildung* Φ_1^{-1} *in das Datenmodell zurück transformiert werden. In der verbundenen Visualisierung kann die so identifizierte Menge* S *hervorgehoben werden.*

Das Konzept des *Linking & Brushing* bietet eine mögliche Perspektive auf das Konzept dieser Arbeit, die hier an dieser Stelle skizziert werden soll: Frameworks, die *Linking & Brushing* unterstützen, verbinden ihre Komponenten auf der Grundlage eines gemeinsamen Datenmodells, über das die Korrespondenz zwischen den einzelnen Visualisierungstechniken hergestellt wird. Dieses Grundprinzip wird für das Konzept dieser Arbeit übernommen und verallgemeinert: Man muss nicht notwendig voraussetzen, dass die Daten nur durch ein einziges Modell beschrieben werden müssen. Man muss ferner nicht voraussetzen, dass es sich bei diesen Modellen stets um Datenmodelle im engeren Sinne handeln muss. *Linking & Brushing* funktioniert dann, solange die Modelle automatisch ineinander transformiert werden können. Data-Mining Verfahren sind Transformationen zwischen Modellen, wobei eines die Daten beschreibt und eines die Muster in den Daten. Das Grundprinzip wird dahingehend verallgemeinert, dass *Linking & Brushing* nicht nur eine Korrespondenz zwischen

verschiedenen Visualisierungen des gleichen Modells herstellen kann, sondern auch die Korrespondenz zwischen den Visualisierungen verschiedener Modelle. Auf diese Weise können die Beschreibungen auf verschiedenen Abstraktionsniveaus der Analyse direkt miteinander gekoppelt werden.

2.4.4 Visuelle Wahrnehmung

Im Rahmen dieser Arbeit wird eine Möglichkeit für die Aufteilung der Aufgaben zwischen Mensch und Maschine vorgeschlagen. Dafür wurde in Abschnitt 2.1 argumentiert, warum diese Aufgaben nicht vollständig von automatischen Methoden übernommen werden können. Unter anderem soll in diesem Abschnitt entsprechend begründet werden, warum diese Aufgaben nicht vollständig durch visuell-interaktive Verfahren umgesetzt werden können und wo die spezifischen Stärken und Schwächen der menschlichen visuellen Wahrnehmung liegen. Das Wissen über den Wahrnehmungsprozess ist natürlich relevant für das Design von Visualisierungstechniken. Tuftes *„above all else, show the data“* [Tuf83] bezieht sich auf den Anspruch, Visualisierung so zu gestalten, dass die in den Daten enthaltenen Informationen möglichst getreu wahrgenommen werden können. Wenngleich der Anspruch nach wie vor besteht, wurde das Primat dieses Anspruchs inzwischen von Amar et al. [AES05] angefochten (siehe Abschnitt 2.4.2).
MacEachren [Mac95, Seite 33ff] gibt einen ausführlichen Überblick über die Entwicklung der Wahrnehmungsmodelle in den vergangenen Jahrzehnten. Nach MacEachren wird Wahrnehmung heute nicht mehr als passiver Prozess der Informationsverarbeitung verstanden, in dem Signale der Sensoren für die kognitive Verarbeitung aufbereitet werden. Praktisch alle modernen Modelle der Wahrnehmung *(ebd., Seite 45)* gehen davon aus, dass Wahrnehmung ein aktiver Vorgang ist, bei dem kognitive Prozesse mit den Informationen von den Sinnesorganen interagieren; ein Prozess, der also insbesondere durch höhere kognitive Prozesse beeinflusst wird.
Auch unter diesem Blickwinkel betrachtet, verschiebt sich die zentrale Anforderung an die Informationsvisualisierung: Die Kernfrage besteht nicht mehr nur darin, zu bestimmen, durch was sich Visualisierungstechniken auszeichnen, die bestimmte Daten getreu darstellen, sondern darin zu bestimmen, durch was sich Techniken auszeichnen, die für bestimmte Aufgaben besonders gut geeignet sind.
Das Modell von Ware [War04b] für den Wahrnehmungsprozess zeichnet im Wesentlichen den Weg der sensorischen Reize im Sehzentrum des Gehirns nach. Dieses soll im Folgenden näher beschrieben werden. Grundlegend für dieses Modelle ist die Unterscheidung zwischen *sensorischen* und *arbiträren* (d.h. beliebigen) Symbolen. Beide unterscheiden sich darin, ob die Verarbeitung und Identifikation der Symbole im Gehirn gelernt werden muss oder nicht. Für die Verarbeitung sensorischer Symbole können dedizierte Regionen in der Sehrinde identifiziert werden, die sich in den ersten Lebensmonaten entwickeln. Ware bezeichnet sie (implizit) als „nicht-beliebig“, weil die Wahrnehmung auf elementarer Ebene bei allen Menschen nahezu gleich funktioniert und weder durch Lernen noch durch andere kognitive Einflüsse beeinflusst wird. Durch die Spezialisierung und (weitgehende) Fixierung der Signalverarbeitungsprozesse im Sehzentrum werden sensorische Symbole schnell verarbeitet. Nach diesem Modell ist mindestens der erste Teil der Verarbeitung der Sinnesreize tatsächlich ein weitgehend pas-

siver Prozess (d.h. der Mensch kann diesen Prozess nur durch über die Blickrichtung aktiv beeinflussen).
Arbiträre Symbole sind dagegen prinzipiell austauschbare Bezeichner für beliebige Konstrukte der Wahrnehmung oder des Denkens. Zeichen und Wörter der Sprache etwa können nicht direkt wahrgenommen, sondern müssen gelesen werden. Die Bilder dieser Zeichen stehen dabei praktisch nie für die Zeichen selbst; deren Bedeutung ist im Allgemeinen kulturelle Konvention. Die Beziehung zwischen Zeichen und Bedeutung muss entsprechend gelernt werden. Die Interpretation der Zeichen ist eine kognitive Leistung, die wegen der Beschränkung des Arbeitsgedächtnisses langsamer und mit niedriger Bandbreite vonstatten geht. Die Unterscheidung zwischen sensorischen und arbiträren Symbolen entspricht, vereinfacht ausgedrückt, der Unterscheidung zwischen visuellen Artefakten und verbalen Artefakten.
Ware stellt durch diese Unterteilung auch die beiden zuwiderlaufenden Anforderungen des Wahrnehmungsprozesses gegenüber. Die erste Anforderung ist die effektive Reduktion von Informationen in den ersten Schritten des Prozesses, die durch fixierte Verarbeitungsschritte im visuellen Kortex erfüllt wird. Die zweite Anforderung ist die flexible, also erlernbare Weiterverarbeitung der bereits reduzierten Informationen. Ware betont dabei, dass diese Unterteilung nur die beiden Extreme der Signalverarbeitung im Gehirn beschreibt. Der Prozess enthält mehrere Zwischenstufen, die jeweils aufeinander aufbauen. Ware fasst den Wahrnehmungsprozess daher in drei Stufen zusammen:
Die erste Stufe beschreibt die Wahrnehmung elementarer visueller Merkmale des Bildes, wie Farbe, Form, Bewegung und andere visuellen Attribute. Die Verarbeitung verläuft dabei massiv parallel, funktioniert weitgehend ohne kognitiven Aufwand, und *präattentiv*, d.h. ohne, dass der Mensch seine Aufmerksamkeit bewusst auf die Signale lenken muss. Als Konsequenz kann der Mensch auf dieser Stufe der Wahrnehmung Informationen finden, nach denen er gar nicht gesucht hat. Eine Visualisierungstechnik, die diese Wahrnehmungsprozesse ausnutzt, bietet den Vorteil, dass das Ergebnis des Prozesses nicht von der momentanen Aufmerksamkeit des Menschen abhängt (vorausgesetzt, dass die Aufmerksamkeit überhaupt auf den Bildschirm gerichtet ist).
Die zweite Stufe ist eine Zwischenstufe für die Segmentierung des wahrgenommenen Bildes. Diese Zwischenstufe kombiniert dabei die Ergebnisse der ersten Stufe des Wahrnehmungsprozesses mit der bewussten Steuerung durch die Aufmerksamkeit des Betrachters. Abgrenzen läßt sich diese Stufe von der ersten Stufe dadurch, dass sie von nachgeordneten Prozessen gesteuert und durch Lernen verbessert werden kann. Die Segmentierung des Bildes in „ungelernte“ Muster erfordert eine langsamere, serielle Verarbeitung der Signale. Muster jedoch, die durch Lernen internalisiert wurden können dagegen auch in der zweiten Stufe präattentiv verarbeitet werden. Die Signale aus der ersten Stufe müssen für die Serialisierung gefiltert werden. Sie ermöglicht ihrerseits aber auch die motorische Reaktion auf diese Signale.
Diese Stufe ist die wichtigste Stufe für die Mustererkennung in der visuellen Wahrnehmung und für das Verständnis dafür, wo die Grenzen für das Design von Visualisierungstechniken liegen. Die Segmentierung des Bildes beruht auf der Integration verschiedener visueller Attribute, die jede für sich potentiell unabhängige Informationsträger sein können. Ein Muster konstituiert einen räumlich abgrenzbaren, nicht notwendig zusammenhängenden Bereich, in dem diese Attribute die gleiche oder visuell ähnliche Ausprägung haben.

Jedoch weist MacEachren [Mac95, Seite 106] unter Berufung auf die Theorie der Merkmalsintegration von Treisman und Gelade [TG80] darauf hin, dass keineswegs alle visuellen Attribute für die Darstellung von Mustern kombiniert werden können. Ware stellt in [War04b, Seite 144ff] etwa zwanzig visuelle Attribute zusammen, die gewissermaßen das Repertoire für das Design einer Visualisierung darstellen. Bis auf wenige Ausnahmen ist es jedoch so, dass die parallel ablaufenden Wahrnehmungsprozesse nicht mehr als zwei bis vier unabhängige Merkmale integrieren. Eine visuelle Suche, die mehr Attribute berücksichtigt, muss sequentiell durchgeführt werden und ist daher auch weniger effizient. Letztlich werden alle visuellen Attribute letztlich aus dem Rasterbild abgeleitet, dass die Netzhaut empfängt und das nur Orts- und Farbinformationen enthält. Für die Ableitung von Farbe, Bewegung, Textur, Stereoskopischer Tiefe und Helligkeit identifiziert Treisman fünf parallele, und daher voneinander unabhängige Signalwege. Er postuliert, dass visuelle Attribute, die ihrerseits eine Kombination dieser fünf Signale sind, nicht mehr parallel integriert werden können. In der Konsequenz bedeutet dies, dass das Repertoire visueller Attribute für die präattentive Erkennung von Mustern wesentlich kleiner ist als das Repertoire für das Design einer Visualisierung insgesamt.
Die sensorischen Symbole, die auch in der zweiten Stufe noch präattentiv wahrgenommen werden, entsprechen jedoch nicht genau den Symbolen, die nicht gelernt werden müssen oder gelernt werden könnten. Ware betont im Gegenteil, dass die präattentive Erkennung von Mustern (als Domäne der zweiten Stufe seines Modells) durch Lernen und Übung verbessert werden kann. Diese Verbesserung kann dazu führen, dass auch hochkomplexe Muster unter Umgehung höherer kognitiver Prozesse identifiziert werden können. Die häufig als Referenz herangezogenen Leistungen von (Blitz-)Schachspielern sind kein Indikator für besondere kognitive Intelligenz, sondern ein Resultat eines solchen Lernprozesses. Die Reaktion auf diese Muster kann daher weitgehend ohne Belastung höherer kognitiver Zentren des Gehirns ablaufen.
Die kognitive Arbeit findet hauptsächlich in der dritten Stufe von Wares Modell statt. Sie beinhaltet höhere kognitive Leistungen wie die Formulierung von Zielen und die Strategien für die Suche, bzw. die Lenkung der Aufmerksamkeit, die wiederum die Wahrnehmung der zweiten Stufe steuert. Darüber hinaus ist sie verbunden mit dem Sprachzentrum des Gehirns. Die Aufgabe der dritten Schicht besteht also darin, zwischen visuellen and verbalen Artefakten des Denkens zu vermitteln. In dieser Stufe erfordert die Aufgabe die vollständige Aufmerksamkeit und Konzentration.
Die Abgrenzung zwischen zweiter und dritter Stufe ist - insbesondere über einen größeren Zeitraum betrachtet - nicht starr. Anderson [And96, Seite 100] bemerkt, wie beim Lernen einer Aufgabe die höheren kognitiven Prozesse hin zu den vorgelagerten Prozessen verschoben werden. Eine Möglichkeit, diesen Effekt zu nutzen, ist aber gerade bei der explorativen visuellen Datenanalyse nur dann gegeben, wenn sich Methodiken und Abhängigkeiten in den Daten über einen längeren Zeitraum nicht ändern. Dies ist jedoch insbesondere bei der aktiven Suche nach neuen Mustern und Zusammenhängen gerade nicht der Fall.
Allerdings lokalisiert Wares Modell verschiedene Lernprozesse in der zweiten und dritten Stufe. Das Resultat des Lernprozesses in der dritten Stufe der Pipeline ist zum einen die Fertigkeit, ein identifiziertes visuelles Muster in verbale Artefakte zu übertragen und zum anderen aus der rein verbalen Beschreibung eines Musters die Korrespondenz zu einer beliebigen Darstellung herzustellen. Keine der beiden Transformationen ist eindeutig: Eine beliebige

Visualisierung kann Muster aus verschiedenen Problemlösungsszenarien auf die gleiche Weise darstellen und umgekehrt kann ein und dasselbe Muster in verschiedenen visuellen Abbildungen unterschiedlich visuell exponiert werden. Da es keine natürliche Transformation zwischen visuellen und verbalen Artefakten gibt und es dafür auch keine fixierten Hirnstrukturen gibt, müssen diese ebenfalls gelernt werden.
In der zweiten Stufe manifestiert sich Gelerntes als Fertigkeit, Muster zu erkennen und gegebenenfalls darauf sogar „automatisch“ zu reagieren. Das Beispiel des Schachspielers zeigt zudem, dass die Komplexität der Muster, die gelernt werden können, sehr viel höher sein kann, als die Muster, die mit Visualisierungstechniken üblicherweise gesucht werden.

Fasst man die Wechselwirkungen zwischen den drei Stufen in Wares Modell zusammen, läßt sich folgendes ableiten:

1. Die Fähigkeit Muster zu erkennen hängt aus physiologischen Gründen ab von den präattentiven Prozessen für die Verarbeitung auf der ersten Stufe.

2. Die Fähigkeit Muster zu interpretieren hängt davon ab, ob die Korrespondenz zwischen arbiträren und nicht-arbiträren Symbolen hergestellt werden kann.

3. Das Herstellen dieser Korrespondenz ist selbst ein kognitiver Prozess, der unter hohem Lernaufwand und bei gleichbleibenden Bedingungen internalisiert werden kann.

Eine Visualisierung, dass die Wahrnehmung insgesamt unterstützt, muss daher Mustererkennung und Musterinterpretation gleichermaßen berücksichtigen. Die Herausforderung dabei ist die beschränkte Verarbeitungskapazität des Arbeitsgedächtnisses in Verbindung mit der Tatsache, dass die zu verarbeitenden Informationen in der explorativen Analyse häufig neu und potentiell komplex sind. Selbst eine Visualisierung, die eine hohe Komplexität getreu wiedergibt, kann daran scheitern, dass der Aufwand für die Interpretation der Informationen zu groß wird. Ware beschreibt in seinem Modell für die Wahrnehmung visueller Informationen, was Visualisierung leisten kann und trägt die Grundlagen dafür zusammen, die berücksichtigt werden müssen, um diese Leistung zu ermöglichen. Allerdings zeichnet es den ganzen Prozess der Wahrnehmung von den elementarsten visuellen Artefakten bis hin zu den höchsten kognitiven Konstrukten und damit vielleicht auch die Grenzen dessen, was *allein* mit Visualisierung erreicht werden kann.
Durch Wares Modell wird deutlich, dass das *Erkennen* und das *Beschreiben* von Mustern durch den Menschen zwei Schritte der Informationsvisualisierungspipeline sind, die anhand der Stufen durchaus voneinander abgrenzbar sind. Darüber erhebt sich die Frage, ob beide Schritte überhaupt in gleicher Weise durch Informationsvisualisierung unterstützt werden können.
Wenn man das Beschreiben von Mustern nicht unterstützen müsste, würde eine Darstellung ohne Legende und Beschriftungen genügen, da kein Zugriff auf verbale Artefakte notwendig ist. Tatsächlich kann das Lesen von Landkarten und die Suche eines kürzesten Weges zwischen zwei markierten Orten deshalb gelingen, wenn die Ikonographie von Landkarten bereits gelernt und verinnerlicht wurde und in allen Karten desselben Themas weitgehend identisch ist.
Der Lernprozess kann nur dann in Gang kommen, wenn ein Nutzer die Gelegenheit bekommt, die Visualisierung und ihre visuellen Symbole zu „lernen“ und mit ihrer - von der

Visualisierung unabhängigen - Bedeutung zu verknüpfen. Im Normalfall geschieht dies durch Legenden und Beschriftungen. Ebenso bedeutend für das Design der Visualisierung selbst, ist daher auch die Umsetzung existierender Konventionen, wie Beschriftungen mit visuellen Artefakten zu verknüpfen sind. Diese Möglichkeit berücksichtigt jedoch nicht den Teil des Wahrnehmungsprozesses, der bis zum Erkennen eines Musters bereits abgelaufen ist - die Segmentierung des Bildes. In der explorativen, visuellen Datenanalyse kann dieser Prozess natürlich nicht durch eine a-priori gegebene Beschriftung oder gar Hervorhebung vorweggenommen werden. Legenden und Beschriftungen können sich nur auf die einzelne Datenelemente oder Wertebereiche beziehen, aus denen sich ein Muster zusammensetzt und diese müssen für die Interpretation des Musters einzeln inspiziert werden.

Grundsätzlich stellt sich die Frage, ob und wann man davon sprechen kann, dass eine Visualisierungstechnik das *Beschreiben* von Mustern unterstützt. Es existiert eine Reihe von Visualisierungs- und Interaktionstechniken, die es ermöglichen, effektiv beliebige Detailinformationen zu den dargestellten Datenelementen zu bekommen, die in der Wahrnehmung zu einem Muster zusammengefasst werden. Jedoch ist zu unterscheiden, nach welcher Strategie der Anwender anschließend vorgehen muss, um die Beschreibung zu erhalten.
Der Strategie, für jedes Datenelement einzeln Detailinformationen abzurufen und auf der

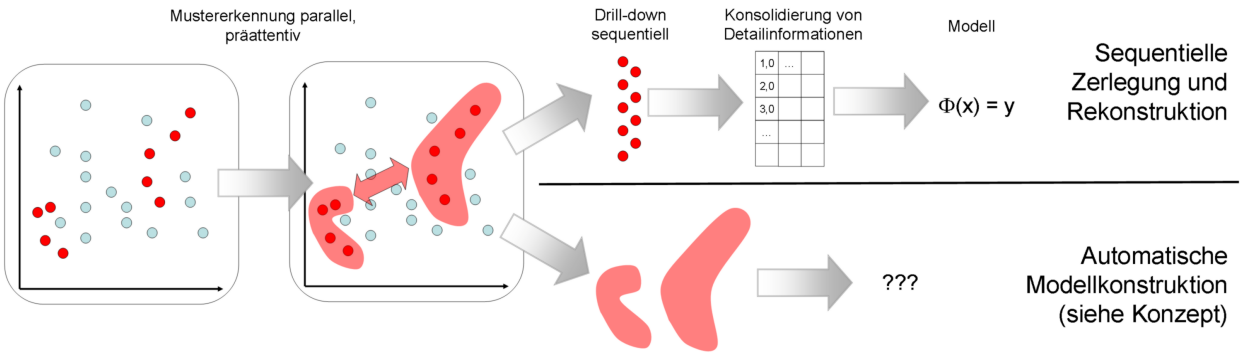


Abbildung 2.16: *Das Beschreiben von Mustern wird von existierenden Visualisierungssystemen bereits häufig in dem Sinne unterstützt, dass über die Interaktion effektiv Detailinformationen abgerufen werden können (oben), die nacheinander zu einem Modell konsolidiert werden können. Dieser Prozess erfordert jedoch eine Zerlegung des bereits erkannten Musters in seine Bestandteile - die Datenobjekte, die zum Muster gehören. Unabhängig davon, wie gut die Detailinformationen durch Interaktion abgerufen werden können, wird mit diesen Verfahren die eigentlich effektiv unterstützte Erkennung von Mustern durch einen kognitiv aufwändigen, sequenziellen Suchprozess für deren Beschreibung fortgesetzt. Um diesen Suchprozess zu umgehen, muss - aus der Sicht des Anwenders - das Muster als Ganzes betrachtet und schließlich verarbeitet werden (unten). Das Konzept dieser Arbeit schlägt eine Strategie dafür vor.*

Basis dieser Detailinformationen eine Konsolidierung und formale Beschreibung des Muster selbst durchzuführen, ist durch das Arbeitsgedächtnis des Menschen enge Grenzen gesetzt, denn dabei handelt es sich um einen sequentiellen Prozess. Selbst wenn eine Technik die Darstellung einzelner Details effizient unterstützt, kann man nicht davon sprechen, dass auch die Konsolidierung und Musterbeschreibung unterstützt würde. Erst dann, wenn diese Aufgaben nicht vollständig durch den Menschen durchgeführt werden müssen, kann man von einer Unterstützung bei dieser Aufgabe sprechen (siehe Abbildung 2.16).
Ware [War04b, Seite 305ff] betont, dass die Konstruktion einer Verbindung zwischen Bild

und Wort (i.e. zwischen sensorischen und arbiträren Symbolen) ein aktiver, kognitiver Prozess ist, der nur unterstützt, jedoch dem Menschen nicht abgenommen werden kann. Er beschreibt dafür mehrere Strategien, die im wesentlichen darauf gründen, zwischen den beiden Codierungen eine unmittelbare Korrespondenz herzustellen. Weiterhin sollen die beiden Codierungen so oft wie möglich gemeinsam genutzt werden. Dies kann besonders in der Datenanalyse dadurch motiviert werden, weil auf diese Weise ein Abgleich zwischen Artefakten auf mehreren Abstraktionsebenen möglich wird.
Erweitert man das technologische Repertoire um automatische Verfahren für die Datenanalyse, besteht nun die Möglichkeit, die Beschreibung eines Musters automatisch durchzuführen: Sie gründet auf der Annahme, dass in erster Linie nur relevant ist, *dass* eine Korrespondenz zwischen Muster und Beschreibung herstellt wird, jedoch nicht vorbestimmt ist, *wie* das geschehen muss. Menschen können Korrespondenzen zwischen Wahrnehmungen herstellen, auch wenn der zugrundeliegende verbindende Mechanismus nicht bekannt ist[6]. Das Modell von Ware liefert damit eine Abgrenzung jener beiden Prozesse - Mustererkennung und Beschreibung - entlang derer im Konzept dieser Arbeit die Schnittstelle für die Arbeitsteilung zwischen Mensch und Maschine etabliert wird.
Im folgenden Kapitel wird ein Modell vorgestellt, in dem die Informationsvisualisierungspipeline kombiniert wird mit dem Wahrnehmungsmodell von Ware. Zusätzlich zu den Verarbeitungsprozessen, die automatisch stattfinden und mit denen das Bild als Medium zwischen Maschine und Mensch erzeugt wird, werden die Prozesse der Mustererkennung und Musterbeschreibung integriert. Ein Schwerpunkt des Konzeptes dieser Arbeit wird auf der Fragestellung beruhen, ob es möglich ist, Analyseprozesse so zu gestalten, dass ein Nutzer genau bei den Teilprozessen involviert ist, bei denen dieser seine spezifischen Stärken hat.
Betrachtet man die drei Teilaufgaben

- Erkennung,
- Beschreibung,
- Interpretation,

dann ist zu untersuchen,

1. wie und unter welchen Bedingungen es möglich ist, diese Aufgaben jeweils durch Visualisierung zu unterstützen,
2. wie es möglich ist, diese Aufgaben jeweils durch Data-Mining Ansätze zu unterstützen,
3. und unter welchen Gesichtspunkten eher die eine oder andere Technologie zum Einsatz kommen muß.

[6]Leider funktioniert das sogar häufig auch in solchen Fällen, in denen ein solcher Mechanismus gar nicht existiert, sondern nur ein Artefakt selektiver Wahrnehmung ist.

2.4.5 Visualisierungstechniken für die Analyse hochdimensionaler Daten

In diesem Abschnitt sollen interaktive Visualisierungstechniken vorgestellt werden, die in erster Linie für die Datenanalyse entwickelt wurden - im Gegensatz zu Designs, die in erster Linie zur Präsentation bekannten Wissens dienen. Taxonomien für die Informationsvisualisierung können die Techniken nach der Art der Daten, die auf die charakteristischen visuellen Elemente eines Datensatzes abgebildet werden (siehe Shneiderman [Shn96]). Eine andere häufig verwendete Einteilung unterscheidet die Visualisierungen nach ihrem visuellen Elementen (nach Keim [Kei00]) oder aber nach den Aufgaben, die eine Visualisierung unterstützt (nach Amar und Stasko [AS05]).
In dieser Arbeit geht es unter anderem um die Frage, inwieweit Visualisierungstechniken für die Analyse und Modellbildung in hochdimensionalen Daten genutzt werden können. Ausgangspunkt für diese Übersicht ist daher eine Datentabelle mit im Prinzip beliebig vielen Attributen für jeden Datensatz. Aus technischen und physiologischen Gründen ist die Anzahl der Attribute, die gleichzeitig sichtbar gemacht werden können, bei jeder Visualisierung beschränkt. Dennoch gibt es natürlich Techniken, mit denen höherdimensionale Bezüge dargestellt werden als mit anderen. Im folgenden soll erläutert werden, nach welchen Kriterien die Techniken hier primär unterschieden werden sollen:
Keim charakterisiert und bewertet die Visualisierungstechniken unter anderem nach der Anzahl unabhängiger Datenattribute, die gleichzeitig in einer Visualisierung gezeigt werden können [Kei00]. Im Rahmen dieses Konzepts soll eine Visualisierung parallel zu automatischen Verfahren für die Suche nach Mustern und Strukturen genutzt werden. Der Mensch übernimmt dabei die Aufgabe der Erkennung dieser Muster. Die gleichzeitige Darstellung mehrerer Attribute ist eine notwendige Voraussetzung dafür, aber sie ist keineswegs hinreichend: Für die Erkennung komplexer Muster müssen visuelle Bezüge zwischen diesen Attributen hergestellt werden. Für die Codierung dieser Bezüge stehen nicht beliebig viele visuelle Attribute zur Verfügung, die zudem nicht alle gleichberechtigt wahrgenommen werden. Ware [War04b] unterscheidet etwa Kombinationen von visuellen Attributen, in denen Muster integriert wahrgenommen werden und Kombinationen, in denen Muster separiert wahrgenommen werden. In letzterem Fall werden, gelenkt von der Aufmerksamkeit, nur Muster des einen oder des anderen Attributes gesehen, aber kein Bezug zwischen den beiden repräsentierten Größen. Die Anzahl unabhängiger Dimensionen im Unterschied zu der Dimensionalität der sichtbaren Bezüge sollen in der folgenden Aufstellung besonders untersucht werden; auch deshalb, weil dieser Unterschied seltener dargestellt wird.

Interaktion erweitert die Möglichkeit für die Exploration von Muster im allgemeinen deutlich, dies betrifft einerseits die Änderung der View-Parameter - um andere Bezüge in den Vordergrund zu rücken - und andererseits die Möglichkeit, mehrere Visualisierungen durch Linking & Brushing miteinander zu verbinden. Dies gilt für praktisch alle Visualisierungen. Daher sollen sie auch danach charakterisiert werden, wie komplex die dargestellten Bezüge mit und ohne Interaktion sein können. Neben der Anzahl der Dimensionen kann eine Visualisierung auch nach dem Detailgrad untersucht werden, in dem die Daten dargestellt werden. Inbesondere lassen sich solche Visualisierungen unterscheiden, die einzelne Werte zeigen und solche, die Verteilungen darstellen. Die Relevanz einzelner Werte hängt ab von

der Aufgabe, für die die Visualisierung eingesetzt wird. Speziell bei der Suche nach Mustern ist die Darstellung der Verteilung von Werten relevanter.

Die Visualisierungstechniken, die für die Analyse hochdimensionaler Daten entwickelt wurden lassen sich in folgende Gruppen einteilen (modifiziert nach Keim [Kei00]). Zu beachten dabei ist, dass Visualisierungstechniken praktisch nie nur eine Visuelle Abbildung in „reiner Form“ beinhalten. Gerade neuere Techniken bauen meist auf Kombinationen älterer Ansätze auf.

- *Geometrische Techniken* in denen die Datenwerte durch geometrische Primitive repräsentiert werden.
- *Projektionstechniken*, die sich direkt aus den automatischen Techniken für die Dimensionsreduktion (siehe Abschnitt 2.3.5.4) ableiten lassen.
- *Pixelbasierte Techniken*, d.h. alle Techniken, in denen jeder einzelner Wert aus der Datentabelle auf einen Pixel abgebildet wird.
- *Hierarchische Techniken* in denen die einzelnen Dimensionen in verschiedene Ebenen einer Hierarchie eingebettet sind.
- *Ikonbasierte Techniken*, die multivariate Kombinationen von Datenwerten auf komplexe Formen/Glyphen abbilden.
- *Tabellen*, orientieren sich beim räumlichen Layout an die übliche Anordnung nach Zeilen (für Datensätze) und Spalten (für Attribute) und lassen sich keiner der anderen Gruppen eindeutig zuordnen.

Keim beschreibt gibt in einem Überblick *(ebd.)* nicht nur eine Kategorisierung von Techniken, sondern vergleicht diese Kategorien bezogen auf verschiedene Charakteristika der Daten, der unterstützten Aufgaben und inherenter Eigenschaften der Visualisierungen selbst. Zu den Visualisierungstechniken, die gleichzeitig die meisten Dimensionen darstellen können zählt Keim geometrische Techniken wie *Scatterplot-Matrizen* und *Parallelkoordinaten*, sowie ikon- und pixel-basierte Techniken. Im Vergleich zu den anderen beiden Gruppen gibt es sehr wenige ikonbasierte Techniken. Neuere Techniken wurden beispielsweise vorgestellt von Chen et al. [CHD+03] oder Nocke et al. [NSS05], dennoch haben sie nie die gleiche Bedeutung erlangt, wie Techniken aus den anderen Gruppen. Auf ikonbasierte Techniken soll hier daher nicht näher eingegangen werden.

2.4.5.1 Geometrische Methoden

Zu den populärsten Techniken für die Visualisierung hochdimensionaler Daten zählen die Parallelkoordinaten und Scatterplotmatrizen, die seit zwei Jahrzehnten immer wieder aufgegriffen und weiterentwickelt werden (siehe z.B. [WAG05, WAG06, BZL+08, Sii00]). Mit Parallelkoordinaten können fünfzig bis einhundert Attribute eines Datensatzes nebenander

dargestellt werden. Direkt sichtbar sind ohne Interaktion dabei nur Relationen zwischen zwei direkt benachbarten parallelen Achsen. Johansson et al. [JFLC08] stellen eine Nutzerstudie mit zwei Varianten von Parallelkoordinaten vor. Einer dieser Varianten sind multirelationale Parallelkoordinaten, mit denen Bezüge von einem zu mehreren Attributen dargestellt werden können. Dabei zeigte sich, dass je mehr Attribute gleichzeitig dargestellt werden, desto weniger robust war die Wahrnehmung im Bezug auf verrauschte Muster.
„Direkt sichtbar“ bezieht sich darauf, dass die Muster ohne jegliche Interaktion identifizierbar sind. Interaktion - insbesondere *Linking & Brushing* (siehe Abschnitt 2.4.3.2) - ist eine Strategie, durch die der Betrachter die Daten in mehreren Visualisierungstechniken oder unter mehreren Blickwinkeln betrachten kann. Durch Linking & Brushing wird dabei die Korrespondenz zwischen selektierten Teilmengen in diesen Darstellungen automatisch hergestellt. Die im Kapitel 4 vorgestellte Visualisierungstechnik ist stattdessen mit dem Ziel entworfen, die Dimensionalität der Muster zu maximieren, die ohne Interaktion wahrnehmbar sind. Unabhängig davon, welche der Strategien umgesetzt werden, riskiert man jedoch, dass die gefundenen Muster so komplex sind, dass sie nicht mehr ohne erheblichen kognitiven Aufwand gelesen werden können. Die Unterstützung der Musterbeschreibung durch automatische Methoden ist ein zentraler Aspekt des Konzepts.
Parallele Koordinaten zeigen daher binäre Relationen, wobei wiederum besonders lineare Abhängigkeiten und auch Cluster in der Visualisierung hervorgehoben werden. Durch Umordnung von Achsen [BTK11] oder Hervorhebung von Clustern [JLJC05] - sowohl durch interaktive als auch automatische Methoden - soll der Nutzwert der Visualisierung verbessert werden. Schließlich wird Brushing genutzt werden, um lokale Abhängigkeiten zwischen nicht benachbarten Attributen zu untersuchen. Bendix stellt mit den Parallel-Sets [Ben06] ein Pendant der Parallelkoordinaten für aggregierte Daten vor, das Mengen und deren Beziehungen anstelle einzelne Datensätze visualisiert.
Auch Scatterplotmatrizen stellen binäre Relationen zwischen Daten dar; Scatterplots für alle Paare von Attributen werden in dem Raster angeordnet. Da ein Scatterplot eine gewisse Mindestgröße benötigt, ist die Anzahl gleichzeitig darstellbarer Attribute jedoch kleiner als bei den Parallelkoordinaten. Weil jedoch alle möglichen Relationen bereits darstellt werden, kommt etwa der Anordnung der Attribute weniger Bedeutung zu. Scatterplot-Matrizen zum Beispiel stellen lediglich alle bivariaten Verteilungen von Datenpunkten über einer gegebenen Anzahl von Attributen des Merkmalsraums dar. Die direkt wahrnehmbaren Muster sind daher immer zweidimensional; es kann höherdimensionale Strukturen des Datensatzes geben, die in jeder dieser Projektionen schwach oder gar nicht sichtbar sind. Cui et al. zeigen [CWR06], wie auch in Scatterplotmatrizen kann die Suche nach höherdimensionalen Strukturen durch Brushing erleichtert werden.
Für Scatterplotmatrizen gibt es mehrere Entsprechungen für Daten auf anderen Aggregationsnivau. Betrachtet man einzelne Scatterplots, kann man, anstatt einzelne Datenpunkte darzustellen, die Dichteverteilung in einem Densityplot darstellen (siehe bspw. Bachtaler [BW08]). Eine Kontingenztabelle leistet das Gleiche für kategorische Daten. Korrelationsmatrizen ordnen Paare von Attributen auf die gleiche Weise an wie Scatterplotmatrizen, reduzieren aber die Relationen auf einen einzigen Wert. Beispiele für ihre Nutzung in einem Analysesystem zeigen Ingram et al. [IMI+10]) und MacEachren et al [MDH+03].

2.4.5.2 Projektionstechniken

Projektionstechniken sind direkt aus den entsprechenden Techniken für die Dimensionsreduktion (siehe Abschnitt 2.3.5.4) abgeleitet. Im Prinzip lassen sich durch diese Verfahren beliebig viele Dimensionen abbilden, allerdings sind einzelne Dimensionen nach der Projektion ohne Hilfsmittel nicht mehr separierbar. Das Ziel der Projektion besteht ja häufig darin, die Ähnlichkeiten der Datenpunkte abzubilden anstelle der Ähnlichkeiten der Attribute. Um Abhängigkeiten zwischen zwei oder mehr Attributen zu finden gibt es zwei Möglichkeiten:
Die erste besteht darin, die duale Projektion durchzuführen. Anstelle der Datensätze werden die Attribute als hochdimensionale Datenpunkte betrachtet und auf die Ebene projiziert. Dieser Ansatz wird etwa bei der Analyse von sogenannten Term-Dokumenten-Matrizen eingesetzt (siehe beispielsweise Kumar [Kum09]), in denen sowohl die Ähnlichkeit der Terme und die der Dokumente für die Analyse relevant sein kann. Durch die duale Projektion können die Abhängigkeiten jedoch nur quantitativ, aber nicht qualitativ untersucht werden.
Die zweite Möglichkeit besteht darin zusätzlich zur Verteilung der Datenpunkte auch die Verteilung der Werte einzelner Attribute darzustellen. Auch so können Abhängigkeiten zwischen den Attributen dargestellt werden, allerdings ist dies kein spezifischer Vorteil von Projektionen, da dies auch mit Scatterplots möglich ist, um Werteverteilungen zu vergleichen.

2.4.5.3 Pixelbasierte Methoden

Pixelbasierte Methoden zeichnen sich dadurch aus, dass die Menge der verwendeten visuellen Attribute auf die beiden Elementaren reduziert wird, d.h. die Position und die Farbe eines Pixels. Jeder Pixel der Visualisierung repräsentiert daher auch einen Datenwert. Dadurch erreicht man im Vergleich zu anderen Visualisierungstechniken eine wesentlich höhere Dichte gleichzeitig darstellbarer Datenwerte. Nach der in [Kei00] gegebenen Definition repräsentiert jeder Pixel höchstens einen Datenwert (d.h. den Wert eines Attributes eines Datenobjektes). Nicht alle hier als pixelbasiert eingeordnete Techniken erfüllen diese strengere Vorraussetzung. Ebenfalls nicht immer streng vorausgesetzt wird, dass jeder Datenwert auf genau einen Pixel abgebildet werden muss. Besonders dann, wenn der Nutzer diese Werte mit einem Zeigegerät auswählen muss, dürfen es auch mehrere Pixel sein.
Die verwendeten visuellen Attribute erfordern die Lösung zweier Designprobleme. Jeder Wertebereich eines Attributes muss dabei auf eine geeignete Farbskala abgebildet werden. Das zweite Problem ist das Arrangement der Pixel bezogen auf die einzelnen Datensätze. Da jeder Pixel nur einen Wert darstellen kann, beschreibt die Methode auch eine Funktion, die einen bestimmten Datensatz auf einen bestimmten Pixel abbildet. Keim et al. diskutieren verschiedene Layouts in [KAK95].
Keim beschreibt in [Kei00] theoretische Grundlagen für pixel-basierte Abbildungen. Die Funktion wird dabei betrachtet als Ergebnis eines Optimierungsproblems. Ein Kriterium dieser Optimierung ist die Erhaltung einer ein-dimensionalen Anordnung von Datensätzen in einem zwei-dimensionalen Display, das zweite Kriterium dieser Optimierung ist die Wahl eines Darstellungsbereiches für ein Attribut, so dass die durchschnittliche Distanz zwischen Pixeln, die zum gleichen Datensatz gehören, minimal wird. Wenn die verschiedenen Datensätze in keiner Ordnung vorliegen ist das erste Kriterium irrelevant, das zweite Kriterium

definiert jedoch immer, wie gut Beziehungen zwischen verschiedenen Attributen der gleichen Datensätze erkannt werden können.
Eine Kombination von pixelbasierten Techniken mit georeferenzierten Daten wurde vorgestellt von Sips et al. [SSKS06]. Die Technik, genannt *PixelMaps*, wird bezüglich der vorliegenden Daten, der Ressourcen möglicher Ausgabegeräte (insb. der Auflösung) und möglicher Aufgaben untersucht. In dieser Technik ist die Funktion, die Datensätze auf Pixel abbildet durch die Geo-Koordinaten vorgegeben. Dabei können die Koordinaten jedoch so verzerrt werden, dass der zur Verfügung stehende Platz optimal genutzt wird indem etwa Leerräume (ohne Daten) verkleinert werden.
Pixelbasierte Methoden reduzieren die Visualisierung im Wesentlichen auf die beiden elementaren visuellen Attribute - Farbe und Position. Der Vorteil dabei ist die im Vergleich zu anderen Techniken hohe Informationsdichte und die Möglichkeit der gleichzeitigen Darstellung mehrerer unabhängiger Dimensionen. Dabei kann man Techniken danach unterscheiden, ob die Techniken einen Wert eines Datensatzes in einem Pixel darstellen oder einen (mehrdimensionalen) Wertebereich in einem Pixel zusammenfassen.

2.4.5.4 Hierarchische Methoden

Die Technik des *Dimensional Stacking* ist eine Variante des rekursiven Layout, das auch bei pixelbasierten Methoden genutzt wird. Im Vergleich zu Pixelbasierten Methoden gibt es jedoch zwei bedeutende Unterschiede: Erstens repräsentiert ein einzelnes visuelles Element keinen Datensatz, sondern Wertebereiche - und damit immer eine Menge von Datensätzen, die in diesen Wertebereich fallen. Zweitens werden beim Dimensional Stacking die Datenattribute nicht räumlich getrennt. Jedes visuelle Element stellt daher stets mehrere Wertebereich dar. So beschreiben beispielsweise Langton et al. [LPWH06] Visualisierungstechniken, in denen mehrere Attribute des Datensatzes im gleichen Display zusammengefaßt werden. Die Farbe eines Pixels ist dann über eine Aggregation über diese Datenobjekte definiert. Auch Wnek [WM94] nutzen diese Art des rekursiven Layout, um ihren Ansatz zu illustieren. Die im Kapitel 4 vorgestellte Visualisierungstechnik ist mit diesen Techniken am engsten verwandt. Im Gegensatz zu den meisten anderen Visualisierungstechniken werden beim *Dimensional Stacking* die relevanten Informationen in Frequenzen wiedergegeben. Pixel, die sich auf ähnliche Wertebereiche beziehen, liegen nicht nur direkt nebeneinander (dann könnte man nur zwei Dimensionen abdecken), sondern auch in definierten Abständen in horizontaler oder vertikaler Richtung. Ein Mensch kann sich wiederholende Strukturen wahrnehmen und faßt diese als Muster zusammen, auch wenn sie nicht unbedingt nahe beieinander liegen. Durch die Verwendung von Frequenzen als visuelle Attribute wird es möglich, wesentlich komplexere, mehrdimensionale Abhängigkeiten zwischen Datensätzen und Attributen zu beschreiben und darzustellen.
Der hier vorgestellte Ansatz geht über die genannten Verfahren insofern hinaus, da die Anzahl der Attribute und eine Diskretisierung für jedes Attribut frei definierbar sind. Dies erhöht die Flexibilität des Verfahrens und ermöglicht die Anpassung an spezielle Muster in dem Datensatz. Allerdings macht es das auch unmöglich zu Lernen, wie ein bestimmtes Muster innerhalb der Visualisierung interpretiert werden muss, denn jede Interaktion verändert gleichzeitig den Referenzrahmen, aus dem sich die Bedeutung der sichtbaren Muster ableitet.

2.4.5.5 Tabellen

Die naheliegenste Abbildung einer hochdimensionalen Datentabelle ist die Darstellung der Werte in der tabellarischen Anordnung. Anstatt die Werte numerisch darzustellen, hat man die Möglichkeit sie farblich oder durch geometrische Primitive auf kleinerem Raum zu visualisieren [RC94]. Mit dieser direkten Abbildung kann man ähnlich wie bei Parallelkoordinaten sehr viele verschiedene Attribute gleichzeitig darstellen. John et al. beschreiben eine Erweiterung [JTS08], in der die Daten zusätzlich aggregiert werden können. Auch hier hilft die reine Anzahl darstellbarer Attribute jedoch nicht dabei, komplexere Abhängigkeiten mit vier oder mehr Dimensionen direkt sichtbar zu machen, um sie zu auffinden zu können. Auch in der tabellarischen Visualisierung müssen diese durch den Anwender schrittweise erschlossen werden, und werden nicht direkt als Muster sichtbar.
Auch wenn mehrere Techniken prinzipiell in der Lage sind, zahlreiche Attribute eines Datensatzes gleichzeitig darzustellen, kann man es immer noch als Herausforderung betrachten Abhängigkeiten von mehr als drei Attributen so sichtbar zu machen, dass sie direkt - ohne Interaktion - als Muster sichtbar sind. Die Überlagerungen mehrerer verschiedener, visueller Attribute stoßen schnell an ihre Grenzen, da sie durch die Wahrnehmung nicht in beliebiger Kombination zu einem Muster zusammengefaßt werden. Eine Ausnahme davon stellt das Dimensional Stacking dar, bei dem die Muster als sich überlagernde Frequenzen wahrgenommen werden. Unter bestimmten Voraussetzungen können dann auch Abhängigkeiten sichtbar gemacht werden, die mehr als drei Attribute umfassen. Im Kapitel 4 wird eine Visualisierung entwickelt, die auf dieser Methode aufbaut. Im Konzept wird dargelegt, dass es einen Unterschied zwischen der Aufgabe gibt ein Muster sichtbar zu machen, und der Aufgabe ein Muster lesbar zu machen. Für die letztere Aufgabe sind folgende Methoden eher geeignet.

2.4.5.6 Methoden zur Modellvisualisierung

Modelle repräsentieren hochdimensionale Abhängigkeiten zwischen Daten. In diesem Sinne können die Visualisierungstechniken, auch als hochdimensionale Visualisierungen verstanden werden. Es geht bei ihrem Design jedoch nicht daum, neue Abhängigkeiten zu finden, sondern bekannte Abhängigkeiten - Zwischenergebnisse der Analyse - so verständlich wie möglich darzustellen. Der Mensch soll mit diesen Techniken sein eigenes Wissen mit diesen Ergebnisse vergleichen können. Die Anforderungen an diese Verfahren richten sich daher stärker an die speziellen Aufgabenstellungen der Nutzer als an die zugrundeliegenden Daten. Methoden für die Modellvisualisierung repräsentieren nicht alle möglichen Muster und Abhängigkeiten; ihr Design ist - noch stärker als die Auswahl von bestimmten Visualisierungstechniken für die Daten - dem „Model-Bias“ unterworfen, d.h. den Vermutungen darüber, welche Arten von Abhängigkeiten in den Daten vorhanden sind. In verschiedenen Fällen, wie zum Beispiel in dem Ansatz von Blanchard et al. [BPKG07], von Ankerst et al. [AEK00] oder von Wilkinson und Friendly [WF09] wird die Visualisierung für ein Data-Mining Verfahren entwickelt, um dessen Ergebnisse bewerten zu können und ggf. zu modifizieren.
In dieser Arbeit wird die These vertreten, dass die beiden Aufgaben - die Suche nach Mustern und die Darstellung ihrer Modelle - durch zwei spezialisierte Techniken besser umgesetzt werden kann als durch eine Technik für beiden Aufgaben. Insbesondere muß sich die

Darstellung der Modelle nicht unbedingt nach den Anforderungen für die visuelle Mustererkennung ausrichten (siehe das Ware-Modell in Abschnitt 2.4.4), sondern kann stärker nach den Nutzeranforderungen und -gewohnheiten richten. Borkin et al. zeigen in [BGP+11] die konsequente Umsetzung der Nutzeranforderung für die Analyse im medizinischen Bereich. Das Design der Visualisierung setzt sich intensiv mit dem „Mental-Model“ der Nutzer und ihren Aufgaben auseinander. Dementsprechend ist das Design spezialisiert auf diese Anwendung. Weitere, allgemeinere Beispiele werden im Abschnitt 3.1.5 des Konzepts beschrieben. Im Rahmen des hier vorgestellten Konzepts schlagen Data-Mining Verfahren die Brücke zwischen Visualisierungstechniken, die die Daten auf unterschiedlichen Abstraktionsstufen und für verschiedene Aufgaben darstellen: Eine Visualisierung für die visuelle Erkennung von Mustern in hochdimensionalen Daten wird gekoppelt mit einer Visualisierung, die das Muster repräsentierende Modell darstellt.

2.5 Visual Analytics

Thomas & Cook definieren Visual Analytics als „Wissenschaft des analytischen Schließens, unterstützt durch interaktive visuelle Schnittstellen“ [TC05]. Im Blickfeld ist dabei der gesamte Prozess der Aufbereitung und Transformation von Rohdaten hin zu entscheidungsrelevanten Kriterien und Informationen; die Definition lässt offen, welche technischen Herausforderungen die Wissenschaft motivieren. Spezifischer ist die Definition von Keim et al. [KAF+08]: Visual Analytics wird definiert als Kombination *automatischer* Analysetechniken mit *interaktiven* Visualisierungen, mit dem Ziel mit den Informationen aus großen und komplexen Datensätzen effektiv Wissen zu schöpfen, dieses zu verfeinern, und schließlich „gute“ Entscheidungen zu treffen.
Die Tatsachen, dass einerseits der gesamte Prozess der Analyse - von den Daten bis zur Entscheidung - und andererseits automatische und visuell-interaktive Verfahren miteinander kombiniert werden sollen, machen Visual Analytics zu einer hochgradig interdisziplinären und integrierenden Disziplin. Die verschiedenen Abstraktionsebenen der analytischen Artefakte beschreiben die Domänen sehr unterschiedlicher Technologien für die automatische Verarbeitung von Informationen (siehe Abbildung 2.2). Einige dieser Technologien sind historisch und technisch miteinander verbunden, wie etwa Machine-Learning und Data-Mining, andere Technologien haben sich auseinander entwickelt. Verschiedene Technologien verfolgen die gleichen Ziele mit unterschiedlichen Methoden (Data-Mining und Informationsvisualisierung), oder auch verschiedene Ziele mit nahezu gleichen Methoden (Informations- und Wissensvisualisierung).
Die Weiterentwicklung der einzelnen Forschungsgebiete ist nur ein Teil der Forschung im Bereich Visual Analytics. Die grundlegendere Motivation besteht darin, komplementäre Stärken zwischen diesen Forschungsgebieten zu identifizieren und diese Stärken durch die Verbindung dieser Forschungsgebiete besser zu nutzen. Bezogen auf den gesamten Analyseprozess setzt sich Visual Analytics in besonderer Weise mit der Frage auseinander, wie Daten-, Informations- und Wissensvisualisierung mit den Techniken für die automatische Datenanalyse verbunden werden kann.

Automatische Verfahren müssen eingesetzt werden, wo die Masse der potentiell relevanten Informationen zu gross ist, um durch den Menschen noch verarbeitet zu werden. Die sich stetig verbesserten Möglichkeiten, Daten zu speichern und abzurufen, erhöhen daher gleichzeitig das Potential für die Nutzung automatischer Methoden in der Datenanalyse. Automatische Verfahren allein verbessern jedoch nicht die Fähigkeiten des Menschen, bessere Entscheidungen zu fällen. Menschen, nicht Computer müssen das Wissen aus Daten schöpfen (nach Few [Few09]).
Interaktive Visualisierungstechniken sind einerseits Analyseverfahren eigenen Rechts, gleichzeitig dienen sie jedoch auch als Schnittstellen für die Sichtung, Interpretation und Bewertung analytischer Artefakte und für die Steuerung und Verbesserung der automatischen Verfahren, die diese Artefakte transformieren. Die Analyse wird zu einem Dialog zwischen Mensch und Maschine. Es ist eine, wenn nicht *die* Herausforderung in Visual Analytics, diesen Dialog so zu gestalten, das er die menschlichen Fähigkeiten, und insbesondere die Art, wie Menschen Informationen verarbeiten, organisch unterstützt.
Dies gilt umso mehr, da die Integration und die Kopplung unterschiedlicher Techniken und

Technologien zwar ein größeres Repertoire an Methoden und Verfahren eröffnet, jedoch auch ein neues Komplexitätsniveau bedeutet. Allein gemessen an den möglichen Optionen wird die Analyse dadurch natürlich aufwändiger.
Es gibt jedoch immer Fälle, in denen Menschen Entscheidungen nicht abgenommen werden können und dürfen. Dies ist insbesondere dann der Fall, wenn automatische Verfahren sich als ungeeignet herausstellen oder präziser: *wenn nicht sichergestellt werden kann*, dass automatische Verfahren belastbare Ergebnisse liefern. Durch Visualisierung muss sich der Nutzer in der Analyse mit den Kriterien für die Entscheidungen auseinandersetzen, für die es keine vorgefertigte, automatisierbare Lösung gibt. Die Herausforderung besteht jedoch darin, den Mensch-Maschine-Dialog im Idealfall so zu gestalten, dass die Entscheidung sich - für den Menschen - aus den präsentierten Informationen direkt ergibt. Das im folgenden Kapitel vorgestellte Konzept beruht im Kern auf einer Annahme: Dass auch eine komplexe Entscheidung - etwa darüber ob ein automatisches Verfahren geeignet ist oder nicht - auf einen einfachen visuellen Abgleich reduziert werden kann.
Visual Analytics ist mehr als nur Visualisierung. Nach van Wijk [vW05] müssen Entwickler neuer Methoden überzeugend darlegen, warum die Informationen, nach denen gesucht wird, nicht ohne den Menschen gefunden und nicht ohne den Eingriff des Menschen bearbeitet werden können. Nach Keim et al. [KAF+08] besteht eine Herausforderung von Visual Analytics gerade in der aktiven Suche nach einem geeigneten automatischen Verfahren und der Identifikation seiner Grenzen. Der Mensch behält die Autorität, aber damit auch die Verantwortung, die Kompetenzen zwischen Mensch und Maschine während der Analyse richtig zu verteilen.
van Wijk hinterfragt ebenso den Wert der Interaktion: Interaktion erhöht die Bandbreite der Optionen, mit denen der Anwender seine Verfahren an Daten, Fragestellungen und Situation anpassen kann, um den größtmöglichen Nutzen daraus zu ziehen. Interaktion verschlechtert aus dem gleichen Grund aber die Vergleichbarkeit der Ergebnisse. Daraus folgt, dass die Entscheidungen, die der Anwender bei der Interaktion trifft, ebenso methodisch kontrolliert werden müssen wie seine automatischen Verfahren.

Visual Analytics bietet durch seine Interdisziplinarität auch eine große Bandbreite verschiedener möglicher Forschungsschwerpunkte. Da sich Visual Analytics ebenso an der Schnittstelle zur anwendungsorientierten Forschung befindet, wird diese Bandbreite noch einmal durch die verschiedenen möglichen Anwendungsgebiete potenziert. Tatsächlich beziehen sich etwa die Hälfte aller Veröffentlichungen (36 von 71) zwischen 2006 und 2008 des VAST *(Visual Analytics Science and Technology)* Symposium schon im Titel auf ein Anwendungsgebiet. Einen Schwerpunkt auf ein bestimmtes Anwendungsgebiet gibt es in der vorliegenden Arbeit jedoch nicht.
Im Rahmen des Konzepts dieser Arbeit werden Techniken der Informationsvisualisierung und der automatischen Datenanalyse (Data-Mining, Machine-Learning und Statistik) als ein Repertoire von Techniken betrachtet, dessen Verfahren potentiell unabhängig von Anwendungsgebieten eingesetzt werden können. Dieses Repertoire ist ebenfalls unabhängig von bestimmten Daten und Fragestellungen, und zwar insofern, dass diese im Allgemeinen nicht eindeutig determinieren, welche Techniken für eine bestimmte Analyse verwendet werden müssen.
Untersucht sollen in dieser Arbeit in erster Linie die Möglichkeiten der Kopplung zwischen visuell-interaktiven und automatischen Verfahren. Um die Kopplung allgemein zu kategori-

sieren, gründet die Untersuchung weniger auf spezifischen Verfahren, als vielmehr auf Verfahrensklassen. Innerhalb des vorgestellten Konzepts wird daher eine verallgemeinernde Perspektive eingenommen. Diese Perspektive wird erzwungen durch die Fragestellung, ob ein automatisches Verfahren für die Analyse geeignet ist, denn es bedeutet auch, das jedes Verfahren zur Disposition stehen muss. Allerdings wird das Konzept dieser Arbeit nicht jede Art der Verbindung abdecken, die technisch umsetzbar wäre. Bei der Arbeitsteilung zwischen Mensch und Maschine muss die Leistung der Maschine den Aufwand für die Interpretation und Bewertung ihrer Ergebnisse amortisieren. Für die Bewertung und Charakterisierung möglicher Verbindungen werden daher beide Aspekte untersucht.

Eine zentrale Rolle haben existierende Modelle für die Systematisierung verschiedener Arbeiten in den relevanten Forschungsbereichen. Dies sind insbesondere das Modell für die Informationsvisualisierung von Card, Mackinlay und Shneiderman (siehe Abschnitt 2.4.2) und entsprechend das Modell für die KDD-Prozess von Fayyad, Shapiro und Smyth (siehe Abschnitt 2.2).
Eine Bewertung und Belastung eines Verfahrens und seiner Parameter ist letztlich nur dann sinnvoll, wenn gleichzeitig gezeigt werden kann, dass überhaupt Optionen zur Disposition stehen. Die genannten Modelle und insbesondere auch der reduktionistische Ansatz für die konzeptionelle Beschreibung von Data-Mining Verfahren (siehe Abschnitt 2.3.4), stellen deutlich heraus, inwiefern sich die verschiedenen Verfahren für die automatische Analyse und die Verfahren für die visuell-interaktive Analyse gleichen, und welche Ansatzpunkte für eine Kopplung existieren. Die im Folgenden beschriebenen Kopplungsmodelle bauen insbesondere darauf auf, dass es sich bei der Informationsvisualisierung als auch beim Knowledge-Discovery interaktive und iterative Prozesse handelt. Die Reihenfolge, in der Teilprozesse ablaufen, ist dadurch schon in den zugrundeliegenden Technologien nicht determiniert.

2.5.1 Kopplungsmodelle für Visual Analytics

Keim et al. [KAS04] beschreiben eine Typologie für die Kopplung im *Visual Data Mining*. Visual Data Mining wird häufig synonym zu Visual Analytics gebraucht, bezieht sich im engeren Sinne jedoch auf die Kopplung zwischen Informationsvisualisierung und Data-Mining. Die existierenden Techniken werden durch die Typologie wie folgt klassifiziert:

Visualisierung als Vorbereitung der automatischen Analyse: In diesem Ansätzen wird die Visualisierung zum Beispiel für die manuelle Datenfilterung und -Selektion, oder auch für die Erkennung von Mustern durch den Menschen eingesetzt. Durch die Visualisierung kontrolliert der Anwender beispielsweise die Parameter, die auf der Grundlage von Daten gesetzt werden müssen.

Visualisierung als Nachbereitung der automatischen Analyse: In diesem Fällen dient die Visualisierung beispielsweise der Präsentation und Interpretation der gefundenen Muster und Trends im Kontext der Daten, oder aber der Visualisierung der Modelle selbst. Die Visualisierung kann auch der Einstieg in eine Feedbackschleife, für die Verfeinerung der

Steuerung sein. Diese Verbindung ist eine Serialisierung der Pipelines des KDD-Prozess und der Informationsvisualisierung (siehe die Abbildungen 2.6 und 2.12). Die meisten Ansätze, die automatische Analyse und Visualisierung miteinander verbinden, implementieren mindestens diese Art der Verbindung. Sie kann dabei sowohl für die Präsentation von Ergebnissen genutzt werden, als auch als Ausgangspunkt für einen neuen Interaktionszyklus, der auf der Interpretation und Evaluation der dargestellten Informationen basiert. dos Santos und Brodlie [dSB04] beschreiben diese Verbindung konzeptionell. Beispiele für Ansätze, in denen die Visualisierung den Ausgangspunkt einer neuen Iteration darstellt beschreiben unter anderem Voinea und Telea [VT07], Turdukolov et al. [TKB07].

Visualisierung als hochintegrierter Prozess innerhalb der Analyse: Bei diesem Typ können die meisten Teilschritte der Analyse durch die Anwender überwacht und direkt gesteuert werden. Durch die Visualisierung werden nicht nur die Parameter und Endergebnisse exponiert, sondern auch alle Zwischenergebnisse. Darüber hinaus kann die dient die Visualisierung nicht nur als Medium der Vermittlung von Informationen von Maschine zu Mensch, sondern auch in umgekehrte Richtung dienen.
Beispielsweise Hao et al. [HDK+07], Yang et al. [YWRH03], Nabney und Maniyar [MN06] beschreiben Systeme in denen der Anwender interessante Muster in der Visualisierung identifiert, und damit die Visualisierung indirekt steuern kann. Diese Muster werden analysiert und das Ergebnis der Analyse beeinflusst die Steuerung der Visualisierung.
Garg et al. [GNRM08] und May und Kohlhammer [MK08a] setzen die Visualisierung für die Modellierung und Exponierung des analytischen Modells ein. All diesen Ansätzen gemeinsam ist, dass die Identifizierung und Beschreibung von Mustern (zumindest teilweise) separiert sind, wobei die visuelle Repräsentierung in ein formales Modell übersetzt wird. Hao et al. [HDK+07] bezeichnen die Daten, die direkt aus der Nutzerinteraktion abgeleitet werden als eine wertvolle Informationsquelle.

Im Sinne der zugrundeliegenden Ideen und Ziele für Visual Analytics, geht die Integration der Visualisierung in jeden Schritt der Analyse als Instrument für Steuerung und Feedback am weitesten. Keim et al. motivieren in [KAS04] diesen Schritt dadurch, dass man durch diese Exponierung gezwungen ist, die Steuerung der automatischen Data-Mining Verfahren an jede Applikation neu anzupassen. Allein durch die starke Integration der Visualisierung wird es möglich zwei zentrale Ansprüche bei der Entwicklung von Visual Analytics Technologie umzusetzen:

1. Methodische und systematische Fehler bei der Auswahl, Anwendung und Steuerung von automatischen Verfahren müssen bei der Analyse so augenfällig werden, dass sie nicht ignoriert werden können - im Idealfall weder von Experten noch von Laien.

2. Aus der Erkennung methodischer oder systematischer Fehler soll der Anwender unmittelbar Optionen erschließen können, mit denen er die Anwendung und Steuerung zielgerichtet verändern kann.

An diesen Ansprüchen mißt sich auch das Konzept dieser Arbeit, das dementsprechend auf der Idee der starken Integration zwischen verschiedenen Komponenten für die Analyse aufbaut.

Der *Visual-Analytics-Prozess* [KAF+08] (siehe Abbildung 2.17) ist ein Kopplungsmodell, an dem verschiedene Kopplungsstrategien lokalisiert werden können. Im Visual-Analytics-Prozess werden die beiden Pipelines für Informationsvisualisierung und den Knowledge-Discovery-Prozess einander gegenübergestellt. Automatische und visuell-interaktive Methoden beschreiben an sich verschiedene Strategien mit denen Wissen aus Daten geschöpft werden soll. Jeder Schritt in diesen beiden Pipelines kann für sich genommen genau so gesteuert werden, wie es beispielsweise die entsprechenden Referenzmodelle vorsehen. Nach diesem Modell wird dieses Wissen entweder direkt aus den Visualisierungen oder über die automatisch erstellten Modelle geschöpft. Charakteristisch für den Visual-Analytics-Prozess ist die Erweiterung durch die direkte Kopplung zwischen automatischer Modellierung und Visualisierung in beide Richtungen. Der Visual-Analytics-Prozess ist ein iterativer Prozess. Jede

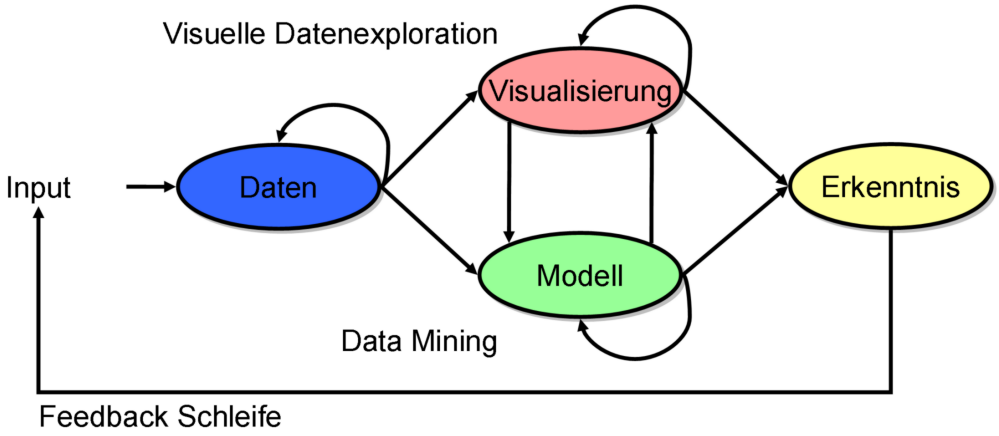


Abbildung 2.17: *Das Modell für den Visual-Analytics-Prozess (nach Keim et al. [KAF+08]). Eine indirekte Kopplung zwischen verschiedenen Teilschritten der Analyse beschreibt die Feedbackschleife. Charakteristisch für Visual Analytics ist jedoch die direkte, enge Kopplung zwischen den Komponenten für die Visualisierung und die automatische Datenanalyse. Wie diese Kopplung umgesetzt wird, kann dabei auf unterschiedliche Weise interpretiert werden (siehe folgende Abbildungen).*

neue Iteration wird über das neu erworbene Wissen des Nutzers definiert, das als Grundlage für Wahl neuer Daten, neuer Verfahren oder deren Steuerung gilt. Betrachtet man den Prozess nur als konzeptionelle Beschreibung einer Iteration, können mit dem Prozess beliebig strukturierte Analyseverfahren allgemein beschrieben werden. Insbesondere läßt sich auch die bereits erwähnte Typologie von Visual Data Mining Verfahren auf den Visual Analytics Prozess abbilden. Der Visual Analytics Prozess ist eine allgemeine, aber insofern auch vereinfachende Darstellung der Kopplung zwischen visuell-interaktiven und automatischen Methoden.
Schneidewind [Sch07] beschreibt für den Visual-Analytics-Prozess bereits mehrere Möglichkeiten der Kopplung. Ein Beispiel dafür ist die Visualisierung von Modellen. In der gleichen Arbeit wird ein Konzept für die umgekehrte Kopplung vorgestellt, in der die Ergebnisse automatischer Methoden dazu verwendet werden, um Visualisierungstechniken zu steuern. Die manuelle Steuerung von Visualisierungstechniken wird ergänzt durch Konzept für die automatische Steuerung, um die Heuristik der Mensch-Maschine Interaktion zu ergänzen mit den automatischen Heuristiken der Data-Mining Verfahren. Diese beiden Varianten werden

unverändert in die hier vorgestellte Systematik übernommen.
Andere Kopplungsvarianten sollen hier auf der Grundlage von Fayyads Modell für das Data-Mining (Abbildung 2.8) und Cards Modell für die Informationsvisualisierung (Abbildung 2.12) verfeinert werden, um unterschiedliche Arten dieser Kopplung beschreiben und einordnen zu können. Im ursprünglichen Modell wurde bereits die Richtung der Beeinflussung unterschieden. Automatische Verfahren können also durch die Visualisierung direkt beeinflusst werden oder umgekehrt.
Im Detail kann man jedoch verschiedene Varianten der Kopplung danach unterscheiden, welche Komponenten eines Verfahrens durch die Kopplung beeinflusst werden. Für die automatischen Verfahren kann man nach dem reduktionistischen Modell (siehe Abschnitt 2.3.4) drei verschiedene Ansatzpunkte für eine Kopplung unterscheiden:

- Verfahrensparameter
- Modellparameter
- Eingabedaten

Wenn durch die Kopplung die Verfahrensparameter verändert werden, wird das automatische Verfahren durch die Visualisierung direkt gesteuert. Die Visualisierung stellt dementsprechend nicht notwendigerweise Daten dar, sondern Verfahrensparameter. Ein Beispiel dafür sind so genannten *Reveiver Operating Characteristics (ROC)* Graphen [BN01], in denen zwei Gütecharakteristika von Klassifikatoren (Sensitivität und Spezifität) gegeneinander abgetragen werden. Im Gegensatz zur Wahl von Verfahrensparametern über die Elemente eines GUI bietet eine Visualisierung die Möglichkeit, den Parameter im Kontext weiterer Faktoren zu untersuchen und zu wählen.
Wenn durch die Kopplung die Modellparameter verändert werden, kann auf diese Weise das Modell manuell editiert werden. Dies kann einerseits in Ergänzung des automatischen Konstruktionsprozess geschehen, der das Modell erzeugt. Es ist im Prinzip jedoch auch möglich die Heuristik des Data-Mining Verfahrens zu ersetzen. Vom Data-Mining Verfahren wird dann nur das Modell verwendet und dieses wird manuell erzeugt.
In dieser Arbeit wird die Visualisierung hingegen als Datenquelle aufgefasst. Mit jeder Visualisierungstechnik, die *Linking & Brushing* unterstützt (siehe Abschnitt 2.4.3.2), ist man auch in der Lage, eine Teilmenge des Datensatzes zu erzeugen. Dies ist eine der beiden Kopplungsvarianten, die im Konzept dieser Arbeit beschrieben werden. Durch die Nutzung des Datenmodells als intermediäre Beschreibung (siehe Abbildung 2.19 (links)) wird eine sehr generische Art der Kopplung zwischen verschiedenen Verfahren möglich.
Beispielsweise nutzen Shrinivasan und van Wijk in [SvW08], aber auch Chen et al. [CBY10] nutzen *Brushing* um interessante Merkmale innerhalb einer Visualisierung zu selektieren. Das Besondere an diesen Techniken ist, daß die so definierten Mengen annotiert werden, um sie zusammenzufassen, um alle Fakten übersichtlich darzustellen und verknüpfen zu können. Dabei wird eine Brücke geschlagen zwischen den Mustern in den Daten und dem Wissen, das bei der Analyse gewonnen wird. Die Beschreibung der Menge kann dabei formale, aber auch natürlichsprachliche Anteile enthalten. Chen et al. schlagen dabei Taxonomien und Templates vor, um die Annotation so weit wie möglich zu automatisieren. Im Unterschied zum hier vorgestellten Konzept werden jedoch keine Data-Mining Verfahren eingesetzt, um

die Muster formal zu beschreiben.
Die größten Ähnlichkeiten zu dieser Arbeit (siehe auch May und Kohlhammer [MK08b]) hat die Arbeit von Garg et al. [GNRM08], die jeweils unabhängig voneinander konzipiert wurden. In diesen Arbeiten werden Mengen durch Brushing definiert, wodurch die Datensätze Klassenlabel erhalten, die für die folgende automatische Modellierung genutzt werden. Garg et al. nutzen dabei induktive logische Programmierung für die Konstruktion von Schlußregeln auf den nutzerdefinierten Mustern. In beiden Fällen hilft ein visuelles Feedback dabei, das Modell zu untersuchen. Ein Unterschied zwischen diesen Arbeiten bestehen in den eingesetzten Techniken für die Visualisierung hochdimensionaler Daten und für die Modellierung. Diese Arbeit geht hinsichtlich des Feedback und der Evaluierung der Modelle noch einen Schritt weiter: Das hier vorgestellte Konzept dient nicht nur der Evaluierung eines bestimmten Modells, sondern zusätzlich auch zur Evaluierung der gewählten Modell*klasse*. Das in Abschnitt 3.2.3.2 vorgestellte „Fuzzy-Feedback“ erlaubt zusätzlich die Darstellung von Modellen auf unterschiedlichen Detailstufen zur Vermeidung des „Overfitting“. In einem neueren Ansatz zeigen Garg et al. [GRM10] wie Interaktion in graph-basierten Daten für diese Kopplung genutzt werden kann. Für die semi-automatische Konstruktion von Graph-Clustern können die Knoten innerhalb des Graphen beliebig innerhalb der Cluster verschoben werden oder aufgeteilt um mehrdeutige Zuordnungen aufzulösen. Da die Cluster dabei direkt verändert werden, handelt es sich um eine explizite (siehe Abbildung 2.9c)) Modellbeschreibung im Unterschied zu den anderen beiden Ansätzen, die das Modell indirekt - als Eingabedaten für ein automatisches Verfahren - verändern.
Für eine Kopplung in die andere Richtung kann man in den Visualisierungstechniken ent-

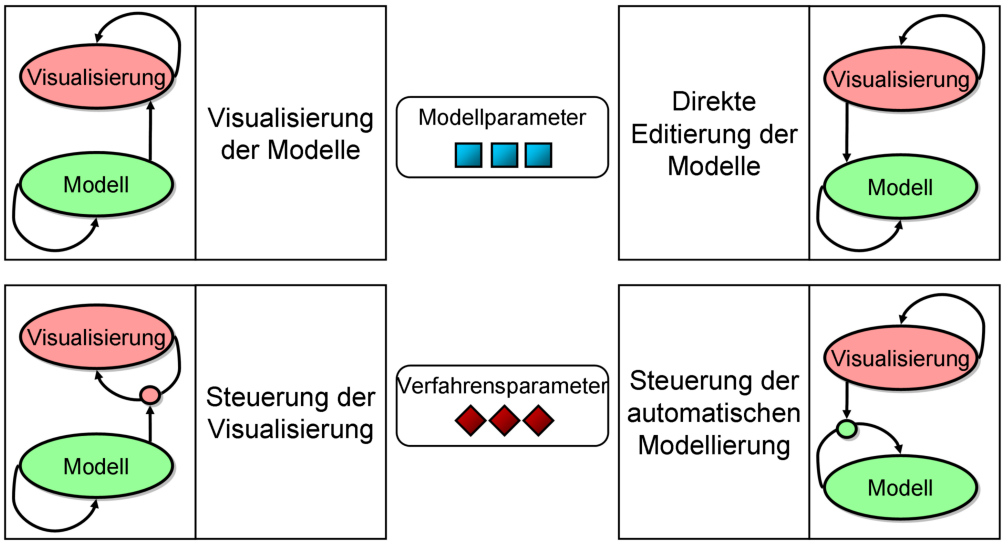


Abbildung 2.18: *Die Kopplung zwischen automatischen und visuell-interaktiven Verfahren kann an verschiedenen Punkten ansetzen. Eine Kopplung über die Modellparameter geschieht einerseits über die manuelle, visuelle Editierung eines Modells, andererseits über die Visualisierung eines automatisch erzeugten Modells. Ebenso können automatische und visuell-interaktive Verfahren über die jeweils komplementären Techniken gesteuert werden. In diesem Fall werden die Verfahrensparameter verändert.*

sprechende Ansatzpunkte finden. Schneidewind beschreibt in [Sch07] zwei verschiedene Varianten: Ebenso wie für automatische Verfahren lassen sich auch für Visualisierungstechniken

Verfahrensparameter identifizieren. Schneidewind beschreibt ein Konzept für die automatische Steuerung von Visualisierungstechniken auf der Basis der Modelle, die wiederum automatisch aus den Daten generiert wurden.
Die Verfahrensparameter einer Visualisierung sind gerade jene, die durch die interaktive Steuerung verändert werden können. Da es, aus der Perspektive der Technik, unerheblich ist, wer oder was die Parameterwerte definiert, lassen sich im Modell von Card et al. (siehe Abschnitt 2.4.2) entsprechend auch verschiedene Ansatzpunkte für die Kopplung ausmachen. Exemplarisch zeigen so Müller et al. [MNS06] wie die Hauptkomponentenanalyse (siehe Abschnitt 2.3.5.4) auf die Steuerung der Datentransformation, der Visuellen Abbildung und der View Transformation angewendet werden kann.
Die zweite Variante ist die Visualisierung des Modells, anstelle etwa einer Visualisierung der Daten. Fasst man die visuellen Attribute einer Visualisierung als Modellparameter auf, dann kann man die Visualisierung eines Modells als Kopplung auffassen, die komplementär zur Editierung eines Modells ist (siehe Abbildung 2.18).

Schließlich besteht die Möglichkeit, ein prädiktives analytisches Modell als Datenquelle aufzufassen. Dies ist die zweite in dieser Arbeit vorgestellten Kopplungsvarianten. Die Kopplung konstruiert einen neuen Datensatz durch eine Simulation auf der Grundlage des Modells. Ziel ist der Abgleich des Modells mit den Originaldaten, der schließlich in der Visualisierung durchgeführt wird.
Berger et al. stellen in [BPFG11] einen Ansatz vor, der Konzepte aus der Informationsvisualisierung und der wissenschaftlichen Visualisierung miteinander verbindet. Für die Berechnung von Leistungsdaten aus Motorparametern wird ein Vorhersagemodell genutzt, das die Abbildung zwischen zwei mehrdimensionalen Räumen beschreibt. Durch eine durch das Modell gekoppelte Visualisierung von (Leistungs-)daten und Modellparametern, lassen sich lokale Charakteristika dieser Abbildung untersuchen. Ziel dieser Analyse ist dabei in erster Linie die Optimierung der Leistungsparameter, jedoch nicht die grundsätzliche Bewertung der Modellfamilie für die Vorhersage (Beispielsweise durch den Abgleich mit Messdaten).
Zusätzlich sind noch zwei weitere Varianten der Kopplung denkbar. Sie unterscheiden sich von den anderen hier genannten Kopplungsvarianten dadurch, das der Ausgangspunkt der Kopplung nicht Ergebnis eines Prozesses ist (d.h. eine Visualisierung oder ein Modell), sondern die Steuerung eines Prozesses. Eine Kopplung kann dadurch umgesetzt werden, dass das Ablaufsprotokoll als Datenquelle aufgefasst wird (siehe Abbildung 2.20). Diese Datenquelle kann wiederum der Ausgangspunkt für (Meta-)Analyse mit Hilfe visuell-interaktiver oder automatischer Methoden sein. Beispielsweise nutzen Shrinivasan und van Wijk [SvW08] eine History-Log für die Visualisierung und die Navigation innerhalb des Analyseprozess.
Ziel dieser Analyse ist nicht allein die Untersuchung der Daten, sondern die Untersuchung und Validierung der Analysemethoden. Diese Kopplungsvarianten sind nicht Gegenstand dieses Konzepts. Allerdings sind die das Gegenstück für jene Kopplungsvarianten, in denen Verfahrensparameter der Ansatzpunkt der Kopplung darstellen. Die Protokollierung früherer Iterationen oder Analyse ist die Grundlage für die Vergleichbarkeit von Ergebnissen und letzten Endes die Grundlage für die Verfeinerung von Analysemethodiken.
Beispielsweise Yang et. al [YXRW07] stellen einen Ansatz für die Konsolidierung der Informationen vor, die innerhalb mehrerer Analysesitzungen aus einem Datensatz geschöpft wurden. Die verschiedenen Ergebnisse können sich ähneln, überschneiden, widersprechen oder in einem anders gearteten Bezug stehen. Das vorgestellte System verbindet eine Daten-

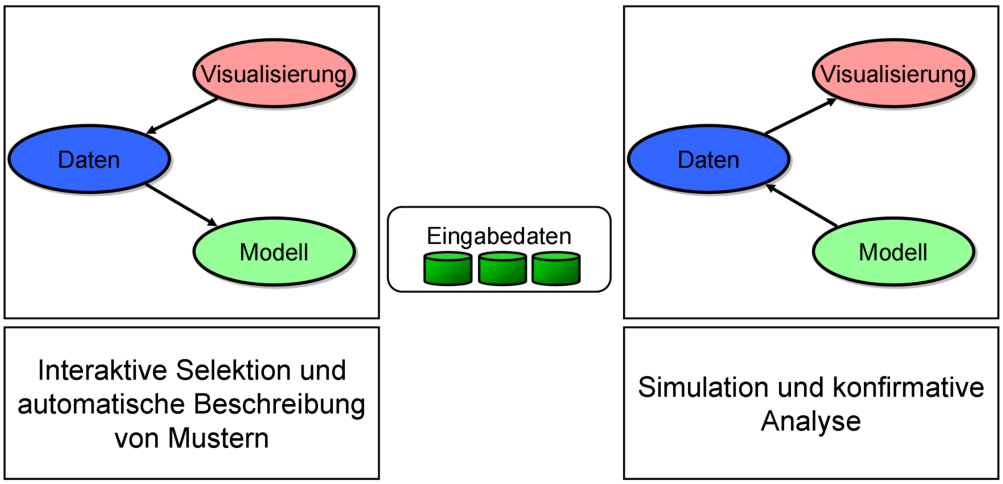


Abbildung 2.19: *Die im Konzept dieser Arbeit vorgestellten Kopplungsvarianten basieren auf dem Prinzip, dass Visualisierungstechniken und automatische Verfahren als Datenquellen für die jeweils komplementäre Technik fungieren können. Wichtig ist dabei, dass im Vergleich zum Visual Analytics Prozess durch die Interpretation der Verfahren als Datenquelle zwei Teilprozesse umgekehrt werden. Es wird die Abbildung einer Visualisierung, bzw. eines Analytischen Modells zurück in den Datenraum etabliert. Im Gesamtkontext der Datenanalyse handelt es sich dabei um die Transformation von Artefakten höherer Stufe (Muster und Modelle) auf die Artefakte der elementaren Stufe.*

visualisierung für die Suche nach Mustern, mit einer Analyse für den Vergleich verschiedener Muster, und die Konsolidierung dieser Ergebnisse.
Einen anderen Ansatz dieser Art beschreiben Rodrigues et al. [RTT05], die eine Visualisierung für den gesamten Analyseprozess vorstellen, wobei Schlussfolgerungen aus Informationen von unterschiedlichem Abstraktionsniveau gezogen werden, die von verschiedenen Ergebnissen stammen. Auch der bereits erwähnte Ansatz von Jankun-Kelly et al. für die Interaktionsanalyse [JKMG07] gehört in diese Kategorie.
Über das Modell für den Visual-Analytics-Process lassen sich alle der vorgestellten Varianten

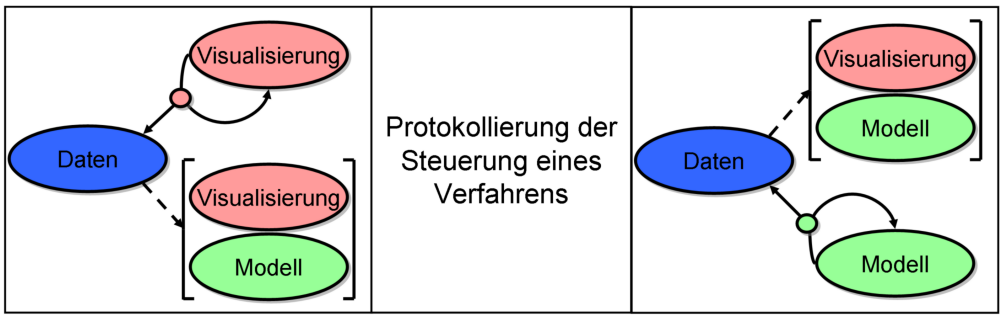


Abbildung 2.20: *Zwei weitere Kopplungsvarianten setzen nicht am Ergebnis der Analyseprozesse, sondern an deren Steuerung an. Die Protokollierung von Verfahren bietet eine Möglichkeit über das Datenmodell automatische Verfahren und Visualisierungen miteinander zu verbinden. Durch diese Verbindung wird gleichzeitig ein Bezug zwischen Analyse und Metaanalyse hergestellt.*

von indirekten Varianten der Kopplung zwischen Visualisierung und automatischer Analy-

se trennen. Indirekte Varianten der Kopplung im Sinne dieser Arbeit sind gerade jene, die über die Feedback-Schleife laufen, wobei in jeder Iteration entweder nur Visualisierung oder automatische Analyse genutzt wird. In diesem Fall stellen die beiden Technologien methodisch zusammenwirkende, aber technisch unabhängige Werkzeuge bereit. Bei der indirekten Kopplung obliegt es dem menschlichen Nutzer, die Kopplung durch die Interpretation und Evaluation der Ergebnisse jeder Iteration in den gemeinsamen Aufgabenkontext herzustellen. Die grundsätzliche Abgrenzung von indirekten Kopplungen und der direkten Kopplungen - welche hier vorgestellt werden sollen - wird im folgenden Konzept detaillierter ausgeführt.
Die verschiedenen Kopplungsvarianten können innerhalb eines Systems für die Analyse mit unterschiedlichen Verfahren eingesetzt werden. Auf diese Weise sind sie frei miteinander kombinierbar. Die Paare der Varianten, die jeweils eine komplementäre Funktion erfüllen, können dabei sogar auf den gleichen Verfahren kombiniert werden: So sind Modellvisualisierung und Modelleditierung kombinierbar, sowie auch die in Abbildung 2.19 vorgestellten Varianten. Dies gilt nicht für die beiden Varianten, in denen sich visuell-interaktive und automatische Verfahren jeweils steuern und auch nicht für die Varianten in denen die Steuerung protokolliert wird. Der Grund dafür ist der, dass der Parameterraum der Verfahrensparameter eines Verfahrens sich im Allgemeinen vom Merkmalsraum der untersuchten Daten unterscheidet.

2.6 Zusammenfassung und Abgrenzung

Das in dieser Arbeit vorgestellte Konzept beschreibt Kopplungsvarianten zwischen Techniken der Informationsvisualisierung und des Data-Mining. Die Aufgaben der Analyse, die durch diese Kopplung unterstützt werden sollen, ist die Suche und Verifikation von Zusammenhängen in komplexen Daten. Im Sinne der Terminologie von Abschnitt 2.1 handelt es sich um die Übersetzung von elementaren analytischen Artefakten über Muster hin zu Modellen und zurück.
Diese Aufgabe soll unter zwei Bedingungen gelöst werden. Einerseits sollen Möglichkeiten für eine Arbeitsteilung zwischen Mensch und Maschine identifiziert werden, die den spezifischen Stärken von Mensch und Maschine gerecht werden. Andererseits soll durch eine Verbindung zwischen explorativer und konfirmativer Datenanalyse eine Möglichkeit etabliert werden, die Übersetzung zwischen analytischen Artefakten zu kontrollieren.
Im vergangenen Kapitel wurde das dafür verwendete Repertoire vorgestellt. Knowledge-Discovery und Data-Mining stellen hierbei das Repertoire automatischer Verfahren. Komplementär stehen die Techniken der Informationsvisualisierung als visuell-interaktive Verfahren gegenüber. Beide Technologien werden für die Übersetzung der gleichen analytischen Artefakte genutzt. Eine Verbindung zwischen diesen Techniken ist dadurch gerechtfertigt, dass sie beide unterschiedliche Vor- und Nachteile besitzen.

Bei der Darstellung beider Technologien wurde besonders Wert gelegt auf die zugrundeliegenden Modelle, die die verfügbaren Techniken möglichst allgemein beschreiben. Die Motivation dafür, eine Kopplung allgemein zu beschreiben, liegt darin begründet, dass eine methodische Analyse erzwingt, dass für die Wahl von Verfahren überhaupt Optionen existieren. Würde die Variante einer Kopplung zwischen visuell-interaktiven und automatischen Verfahren die Wahl der dafür einsetzbaren Techniken determinieren, hätte diese Variante einen geringen praktischen Nutzen.
Das vorgestellte Konzept ist daher auch nicht an bestimmte Techniken gebunden, sondern setzt auf Ansatzpunkten für die Kopplung zwischen den Technologien auf, die möglichst unabhängig von bestimmten Techniken identifizierbar sein sollen. Um diese Ansatzpunkte zu identifizieren, wurden daher bewusst Modelle gewählt, die die unterschiedlichen Techniken ihres jeweiligen Bereiches möglichst umfassend beschreiben.
Für die automatischen Verfahren wurde das reduktionistische Modell von Fayyad et al. (siehe Abschnitt 2.3.4) vorgestellt, das Verfahren des Data-Mining allgemein beschreibt. Durch die Zerlegung von Verfahren in Modelle, Heuristiken und Gütefunktionen können nicht nur die verschiedenen Ansatzpunkte für eine Kopplung identifiziert werden. Dieses Modell ermöglicht auch die Einordnung verschiedener verwandter Analysetechniken hinsichtlich gleicher oder unterschiedlicher Komponenten. Die Vielfalt und die Beziehungen zwischen Verfahren wurde am Beispiel der Entscheidungsbäume dargestellt.
Durch das reduktionistische Modell lassen sich die Freiheitsgrade bei der Verwendung eines bestimmten Verfahrens identifizieren. Das Modell für den Knowledge-Discovery Prozess (siehe Abschnitt 2.2) zeigt die Freiheitsgrade innerhalb des gesamten Prozesses. Die zahlreichen Freiheitsgrade zeigen gerade die Optionen, durch die methodisches Arbeiten erst möglich wird, sie verdeutlichen aber auch ein zentrales Problem der Datenanalyse: Ziel der Datenanalyse ist nicht nur ein Ergebnis. Sie muss gleichzeitig sicherstellen, dass die Ergeb-

nisse nicht ein Artefakt unerkannter oder willkürlicher Entscheidungen oder Annahmen sind, denn in diesem Fall hätten sie eine geringe Aussagekraft.
In den Abschnitten 2.3.2 und 2.3.3 wurde dargelegt, wie die Wahl der Verfahren, die Ergebnisse der Analyse - wie etwa die Komplexität und Lesbarkeit der Modell beeinflussen kann. Die Taxonomie von Han et al. zeigt überdies, dass nicht nur die „klassischen“ Aufgaben des Data-Mining wie Clustering, Klassifikation und Ausreißeranalyse einen bedeutenden Einfluss haben, sondern in gleicher Weise auch die Schritte, die das Data-Mining „nur“ vorbereiten.
Ein Ergebnis - etwa in Form eines Modells der Zusammenhänge in den Daten - repräsentiert jedoch keine Information über seine eigene Aussagekraft. Im Gegenteil, das Modell verwebt Zusammenhänge in den Daten mit allen Entscheidungen, die in der Analyse zu diesem Modell geführt haben. Methodisches Vorgehen beruht stets darauf, sicherzustellen, dass (Teil-) Ergebnisse vergleichbar bleiben. Dazu ist es letztlich nicht nur notwendig, Zusammenhänge in den Daten zu exponieren, sondern auch die Zusammenhänge zwischen Ergebnissen und Parametern der Analyse.
Dies erhöht die Komplexität der Analyse erheblich. Die genannten Modelle zeigen, dass die Anzahl der Freiheitsgrade, mit denen eine Analyse gestaltet werden kann, die Anzahl der Dimensionen des Merkmalsraums, der eigentlich untersucht werden soll, bei weitem übertreffen kann. Um „gute“ Entscheidungen für eine systematische Analyse zu treffen, muss der Analyst eine während der Analyse wachsende Informationsmenge überschauen. Betrachtet man das Konzept aus der Perspektive automatischer Verfahren, werden Techniken der Informationsvisualisierung als Hilfsmittel genutzt, um die Kriterien und Informationen, auf deren Grundlage ein Verfahren gesteuert wird, zu konsolidieren und sichtbar zu machen.
Von den zahlreichen Fragestellungen, die sich bei der Anwendung automatischer Verfahren ergeben, werden in Konzept und Realisierung zwei herausgestellt. Die erste behandelt die Frage, ob das für eine Analyse gewählte Verfahren überhaupt ein „gutes“ Modell für die Strukturen und Muster in den Daten bereitstellen kann. Die Kopplung etabliert eine direkte Kooperation zwischen Mensch und Maschine. Dabei steuert der Mensch das Verfahren dadurch, dass er definiert, welche Datenmenge ein Muster repräsentiert, und erhält direktes Feedback darüber, wie das Muster im Modell beschrieben wurde.
Die zweite Fragestellung behandelt, wie eine Kopplung gestaltet werden könnte, in der innerhalb einer Visualisierung Verfahrensparameter eines automatischen Verfahrens gewählt werden. Exemplarisch wird eine Kopplung für die Wahl von Attributen eines Datensatzes für die weitere Analyse vorgeschlagen.

Das Modell für den Knowledge-Discovery-Prozess hat seine Entsprechung im Modell von Card et al. für die Informationsvisualisierung (siehe Abschnitt 2.4.2). Die Ansatzpunkte für die Interaktion in den einzelnen Schritten des Visualisierungsprozesses stellen in gleicher Weise Ansatzpunkte für die Kopplung dar. Technisch ist es immer möglich, die Parameter eines Verfahrens, die interaktiv gesteuert werden können, auch automatisch zu steuern. Dabei handelt es sich in gewisser Weise um die umgekehrte Richtung der Kopplung zwischen visuell-interaktiven und automatischen Verfahren. Mit dem Modell von Card et al. wurde die Systematik der Kopplungen vervollständigt, anhand derer das Konzept dieser Arbeit in den Forschungsbereich Visual Analytics eingeordnet werden kann.
Von zentraler Bedeutung für diese Arbeit sind ebenfalls die Interaktionsmodelle von Norman und dessen Spezialisierung von Lam (siehe Abschnitt 2.4.3) und die Modelle für die Wahrnehmung als Grundlage der Informationsvisualisierung (siehe Abschnitt 2.4.4). Ge-

meinsam liefern beide den konzeptionellen Rahmen für die Trennung von Mustererkennung und Musterbeschreibung, die die Kopplung von Visualisierung und automatischer Analyse begründet.
Das dreistufige Wahrnehmungsmodell von Ware beschreibt die Prozesse, die bei der Wahrnehmung von Mustern ablaufen und es begründet, inwiefern das Design einer Visualisierung die Fähigkeit des Menschen beeinflusst, in einer Visualisierung Muster zu erkennen. Die letzte Stufe in Wares Modell beschreibt dabei die Transformation von sensorischen in arbiträre - also sprachliche - Symbole und umgekehrt.
Die Interpretation der Wahrnehmung in Lams Interaktionsmodell entspricht weitgehend dieser Transformation. In beiden Fällen bedeutet es, dass die Information, die das Muster repräsentiert, aus dem Kontext des Bildes herausgelöst wird und in den Kontext der Aufgabenstellung übertragen wird. Ab einer gewissen Komplexität der Muster und Zusammenhänge kann der Anwender diese Korrespondenz jedoch nicht mehr selbst ohne Hilfsmittel herstellen.
In der explorativen Datenanalyse kommt das Problem hinzu, dass von vornherein nicht bekannt ist, welche Datenelemente zu einem Muster zusammengefasst werden können. Das bedeutet, dass beispielsweise nicht festgelegt werden kann, welche Strukturen in den Daten eine verbale Beschreibung rechtfertigen oder nicht. Hier setzt die Kopplung mit den Data-Mining Verfahren an, denn diese stellen die Heuristiken und Modelle bereit, um die Beschreibung eines Musters aus den visuelle Artefakten zu erzeugen. Welche Bereiche des Bildes jeweils zu einem Muster zusammengefasst werden, bestimmt dabei der Anwender selbst.
Aus der Sicht der Informationsvisualisierung ist die Kopplung daher eine Hilfestellung für die Herstellung der Korrespondenz zwischen Interaktion und Aufgabe (nach Lams Modell) bzw. zwischen sensorischen und arbiträren Symbolen (nach Wares Modell) durch den Anwender. Der Anwender wird dabei von der Aufgabe entlastet, die innerhalb des Wahrnehmungsprozesses die Komplexität der Informationen, die nicht nur dargestellt, sondern auch genutzt werden können, am stärksten einschränkt.
Dies erhöht das Potential von Visualisierungstechniken, weil es dadurch sinnvoll wird, auch potentielle Zusammenhänge darzustellen, die zu komplex sind, als dass sie unmittelbar interpretiert werden können. Das Design von Visualisierungstechniken muss sich dann weniger an den Beschränkungen des Arbeitsgedächtnis orientieren, sondern kann sich nach der Komplexität von Aufgabe und Daten richten.

Der letzte Abschnitt dieses Kapitels 2.5 lieferte einen Überblick über die Kopplungen zwischen Informationsvisualisierung und Data-Mining, die bereits beschrieben wurden. Grundlage der Systematik sind die Kopplungsmodelle, die in der Literatur im Bereich *Visual Analytics* bzw. *Visual Data Mining* bereits vorgestellt wurden - insbesondere das Modell vom *Visual Analytics Prozess*. Ebenfalls Grundlage sind die Modelle von Fayyad et al. und Card et al. für die Techniken und Verfahren der jeweiligen Technologien.
Beide Grundlagen wurden kombiniert, um eine feinere Systematik der möglichen Kopplungen zu beschreiben. Dahinter steht die Annahme, dass jeder Ansatzpunkt für die Steuerung eines beliebigen Verfahrens sowohl mit Techniken der Informationsvisualisierung als auch mit Techniken der automatischen Datenanalyse verknüpft werden kann. Anhand dieser Ansatzpunkte wurden in der Systematik mehrere Varianten unterschieden. Einige dieser Varianten wurden bereits existierenden Techniken zugeordnet, und andere wurden entsprechend für die Einordnung des hier vorgestellten Konzepts in die Systematik verwendet.

Im Gegensatz zu den meisten existierenden Ansätzen für eine Kopplung zwischen visuell-interaktiven und automatischen Verfahren werden im vorgestellten Konzept Visualisierung und Interaktion als Datenquelle betrachtet. Die Visualisierung ist also nicht in erster Linie nur das Medium, über das die (Zwischen-)Ergebnisse der Analyse an den Nutzer übermittelt werden, sondern in erster Linie das Medium für die Steuerung der Verfahren.
Ferner betont das Konzept die Notwendigkeit einer direkten Korrespondenz zwischen wahrgenommenen Mustern und einer damit gekoppelten Interaktion. Dies rührt von dem Anspruch, Möglichkeiten der Steuerung automatischer Verfahren zu identifizieren, in denen der Anwender dieses Korrespondenzproblem nicht selbst lösen muss. Dies bedeutet nicht, dass dieser Anspruch für jede Teilaufgabe der Analyse befriedigt werden könnte. Das Konzept fokussiert auf die Suche und Beschreibung von Mustern, was nur eine, wenngleich elementare Teilaufgabe der Analyse ist.
Dennoch wird das Repertoire der Operationen, mit denen der Mensch zu der Erkennung und Beschreibung von Zusammenhängen in komplexen Daten beiträgt, bewusst zunächst auf jene reduziert, die wenig oder keinen kognitiven Aufwand erfordern - d.h. das Erkennen und Selektieren von Mustern und Vergleichen von Bildern. Im Konzept wird untersucht, welche Anpassungen von Techniken und Kopplung diese Beschränkung erfordert. Die Ansatzpunkte für die Steuerung automatischer Verfahren, die im reduktionistischen Modell von Fayyad et al. definiert werden können, geben hierbei die Varianten für die Kopplung vor.

Kapitel 3

Konzept

Im letzten Kapitel wurden die Forschungsgebiete beschrieben, auf denen das folgende Konzept aufbaut. Auf der Grundlage des Visual-Analytics-Prozesses wurden acht verschiedene Varianten für die Kopplung visuell-interaktiver und automatischer Verfahren identifiziert (siehe Abschnitt 2.5.1). Die Varianten unterscheiden sich darin, welche Ansatzpunkte eines Verfahrens für die Kopplung gewählt werden. Zwei dieser Varianten sollen im Rahmen dieses Konzepts vorgestellt werden: In der ersten Variante (siehe Abschnitt 3.1) wird die Interaktion in der Visualisierung als Datenquelle für die Kopplung aufgefasst. Jede Visualisierungstechnik kann in dem Sinne als Datenquelle identifiziert werden, da es möglich ist, eine bestimmte Teilmenge der dargestellten Daten durch manuelle Selektion auszuzeichnen. Um eine Datenquelle handelt es sich deshalb, weil die Selektion durch den Nutzer ein neues, binäres Merkmal darstellt, das nicht notwendig vor der Analyse zur Verfügung stand.
Die gemeinsame und gleichzeitige Auswahl beliebiger Daten identifiziert mindestens zwei Mengen. Im Folgenden wird stets davon ausgegangen, dass die Auswahl dieser Mengen durch den Menschen nicht chaotisch erfolgt, sondern einem bestimmten Muster folgt. Im Besonderen sind dabei genau jene Muster und Strukturen gemeint, die der Mensch in der Visualisierung als solche wahrnimmt.
In den Modellen der Informationsvisualisierung ist vorgesehen (siehe Abschnitte 2.4.2 und 2.4.3), dass der Mensch wahrgenommene Muster *selbst* interpretieren muss, um schließlich zu neuen Erkenntnissen zu gelangen. Stattdessen wird hier der „Umweg“ vorgeschlagen, in dem Muster zunächst in das Datenmodell zurück transformiert werden, um anschließend durch automatische Verfahren modelliert zu werden (siehe Abbildung 3.1). Bei diesem Umweg findet eine Arbeitsteilung zwischen Mensch und Maschine statt. Hintergrund ist die Feststellung, dass Mustererkennung und Musterbeschreibung verschiedene Teilprozesse sind. Betrachtet man ausschließlich automatische Suchverfahren, sind diese Prozesse nicht zu unterscheiden - nur ein Muster, das innerhalb der Modellfamilie eines Verfahrens formal beschrieben werden kann, wird auch gefunden.
Der Grund für diese Unterscheidung liegt in der Nutzung von Visualisierungstechniken. Anhand der Modelle von Lam [Lam08] (siehe Abschnitt 2.4.3) und auch von Ware [War04b] (siehe Abschnitt 2.4.4) können Mustererkennung und -beschreibung voneinander abgegrenzt werden. Betrachtet man Techniken der Informationsvisualisierung isoliert, dann ist gemäß dieser Modelle sowohl die Erkennung als auch die Beschreibung von Mustern Aufgabe des

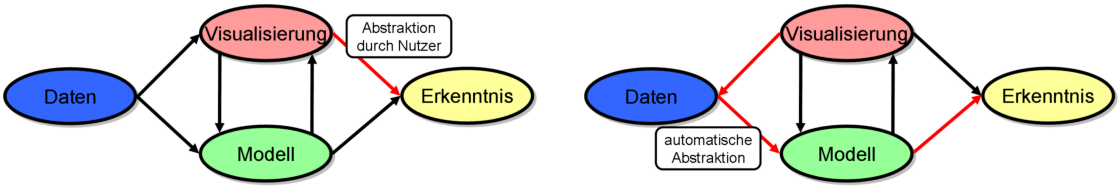


Abbildung 3.1: *Um ein wahrgenommenes Muster in eine Erkenntnis umzuwandeln, müssen die Wahrnehmungseindrücke interpretiert werden. In der Informationsvisualisierung führt der Mensch diese Abstraktion vom wahrgenommenen Bild zur Beschreibung des Musters selbst durch. Im Rahmen dieses Konzepts wird vorgeschlagen, diese Beschreibung durch automatische Verfahren durchzuführen. Dadurch wird eine Aufteilung zwischen Mustererkennung und Musterbeschreibung zwischen Mensch und Maschine etabliert, die den Fähigkeiten beider eher gerecht wird.*

Anwenders. Daraus folgt, dass sich die Gesamtkomplexität der Aufgabe daran bemisst, wie leicht die schwerere der beiden Aufgaben bewältigt werden kann. Selbst wenn eine Visualisierungstechnik genau eine dieser beiden Aufgaben sehr gut unterstützt, kann der Nutzwert dieser Technik allein dennoch gering sein.
Die diesem Konzept zugrundeliegende Annahme besteht darin, dass die beiden Aufgaben nicht notwendigerweise durch die gleiche Technologie unterstützt bzw. durchgeführt werden müssen. Mit der Kopplung automatischer und visuell-interaktiver Techniken steht ein größeres Repertoire an Verfahren zur Verfügung. Die Unterscheidung der beiden Teilprozesse Mustererkennung und Musterbeschreibung erfolgt auf der Grundlage der Forderung, dass die Arbeitsteilung zwischen Mensch und Maschine gerade so gestaltet wird, dass die Fähigkeiten, in denen Mensch und Maschine sich am deutlichsten unterscheiden, auch entsprechend genutzt werden.
Ursprünglich motiviert ist diese Trennung durch eine Fragestellung für die Informationsvisualisierung hochdimensionaler Daten. Angenommen, die Komplexität eines Musters, eines Modells oder einer Aussage sei definiert durch die Anzahl der Bezüge zwischen den Variablen eines Datensatzes. Bestimmt man diese Anzahl, kann man die Komplexität der Muster, die visualisierbar und wahrnehmbar sind, vergleichen mit der Komplexität der Modelle, die der Nutzer ohne Hilfsmittel aus den wahrgenommenen Mustern beschreiben und interpretieren kann. Durch die Beschränkung des Arbeitsgedächtnisses auf wenige Inhalte ist die Anzahl der Bezüge, die gleichzeitig verarbeitet werden können, stärker eingeschränkt als die Komplexität der Muster, die wahrgenommen werden können.
Nimmt man für eine Visualisierung in Anspruch, *beide* Aufgaben (Mustererkennung und Beschreibung) in gleichem Maße zu unterstützen, ergäbe sich der Nutzwert der Visualisierung insgesamt aus der für den Anwender schwierigeren Aufgabenstellung. Konkret wäre es dann beispielsweise nicht sinnvoll, eine Visualisierung zu konstruieren, mit der hochkomplexe Muster erkannt werden können, wenn der Anwender nicht in der Lage wäre, die dahinter stehenden Regeln zu identifizieren.
Das hier vorgestellte Konzept verfolgt diesen Anspruch jedoch ausdrücklich nicht. Die Transformation von Mustern in Modelle wird durch automatische Verfahren erfolgen. Dadurch wird es möglich, das Design einer Visualisierung stärker danach auszurichten, welche Muster der Mensch erkennen kann. Der kognitive „Flaschenhals“ des Arbeitsgedächtnis wird entlastet.

Die zweite Verfeinerung beschreibt die Integration von explorativer und konfirmativer Analyse innerhalb eines Prozesses (siehe Abbildung 3.2). Das Ziel ist ein paralleler Abgleich zwischen analytischen Artefakten auf zwei verschiedenen Abstraktionsebenen. Auf der einen Seite sind dies Daten und Muster, auf der anderen Seite handelt es sich um formale analytische Modelle. Der parallele Abgleich zielt dabei nicht nur darauf ab, effizient eine Korrespondenz zwischen Mustern und Modellen herzustellen. Er dient insbesondere dazu, die Verfahren, die die beiden Abstraktionsstufen ineinander überführen, auf den Prüfstand zu stellen.
Die enge Integration von explorativer und konfirmativer Analyse stellt zwei technisch unabhängige Prozesse - Modellierung und Simulation - einander gegenüber. Für die Analyse ist es eine notwendige Forderung, dass die beiden Prozesse komplementär sind, dass sie somit ihrer Hintereinanderausführung kein vom Ausgangspunkt wesentlich abweichendes Ergebnis liefern. Während die Transformationen selbst automatische Prozesse sind, dient die Visualisierung der Daten und der Modelle dem visuellen Abgleich der Artefakte auf beiden Abstraktionsebenen.
Wie im Abschnitt 2.1 motiviert, besteht das Risiko der explorativen Datenanalyse nicht nur

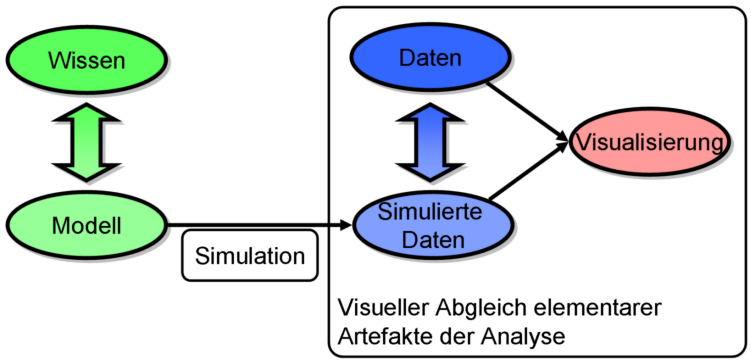


Abbildung 3.2: *Über ein prädiktives analytisches Modell kann durch die Simulation ein Datensatz erzeugt werden, der mit Referenzdaten verglichen werden kann. Der visuelle Abgleich der Datensätze erlaubt eine qualitative Bewertung des Modells, mit der insbesondere unterschieden werden kann, ob das Modell die Muster und Strukturen der Referenzdaten reproduziert, was mit statistischen Kennzahlen alleine nicht funktioniert. Die Kopplung beider Verfeinerungen für die integrierte Konstruktion und Bewertung des Modells kann genutzt werden, um nicht nur die Modelle, sondern auch die Verfahren einzuschätzen.*

allgemein darin, wegen unsicherer Referenzdaten potentiell falsche Ergebnisse zu erhalten. Das Risiko verbirgt sich in der Auswahl von Verfahren und Parametern selbst, mit denen diese Ergebnisse geschöpft werden. Wenn aus den Ergebnissen nicht mehr erschlossen werden kann, ob sie ein Artefakt der Daten, oder ein Artefakt der Auswahl von Methoden und Parametern sind, sind sie nicht aussagekräftig. Insgesamt behandelt die Konzept daher auch die Frage, wie dieses Risiko minimiert werden kann.
Hinter dieser Verallgemeinerung steht die These, dass sich eine Entscheidung innerhalb der Analyse (d.h. für ein Verfahren, für eine Auswahl von Daten, etc.) nicht prinzipiell unterscheidet von Entscheidungen außerhalb der Analyse. „Außerhalb“ bezeichnet hier eine Fragestellung der Anwendungsdomäne, die beispielsweise auf der Basis der Ergebnisse einer Analyse erst beleuchtet werden soll.

Sowohl innerhalb als auch außerhalb der Datenanalyse können Entscheidungen auf unterschiedlichen Grundlagen getroffen werden:

- Es gibt eine Regel, anhand derer die Entscheidung getroffen werden kann.
- Die Informationen und das Wissen, die eine Entscheidung stützen, liegen dem Anwender bereits vor. Die Entscheidungskriterien sind jedoch nicht bekannt.
- Dem Anwender liegen nur Informationen vor, die potentiell mit der Entscheidung in Beziehung stehen, die genaue Beziehung ist jedoch nicht bekannt.
- Der Anwender muss die Informationen, die potentiell relevant sind, erst suchen.

Von oben nach unten erhöht sich der subjektive Grad der Unsicherheit bei der Entscheidung. Objektiv betrachtet kann beispielsweise natürlich auch eine vorgegebene Regel fehlerhaft sein. Der Unterschied zwischen einem „normalen“ Anwendungsszenario und den Entscheidungen in der Datenanalyse besteht darin, dass bei der Datenanalyse potentiell falsche Entscheidungen getroffen werden *dürfen*. Solange sie erkannt werden, bevor Einfluss auf präsentierte Ergebnisse haben, richten sie (abgesehen von Aufwänden) keinen Schaden an. Es gibt jedoch keinen Grund dafür anzunehmen, dass eine Entscheidung innerhalb der Analyse richtig wäre, nur weil die Analyse überhaupt ein Ergebnis geliefert hat.
Ein wichtiger Grund dafür, Visualisierungstechniken in der Datenanalyse zu nutzen, besteht darin, dass Visualisierung es ermöglicht potentiell relevante Informationen zu finden, die man vorher nicht explizit formulieren konnte. Diese Möglichkeiten bieten sich nicht nur für die Identifizierung von Strukturen und Mustern in Daten. Ebenso sinnvoll ist die Anwendung der Visualisierung bei der Suche nach Fehlern in Modellen oder Methoden. Der visuelle Abgleich von Referenzdaten und simulierten Daten erlaubt, im Gegensatz zu rein statistischen Kennzahlen, qualitative Informationen über die Fehler. Systematische Fehler in der Analyse können sich beim visuellen Abgleich als Muster manifestieren. Auf diese Weise können die Methoden der explorativen Analyse auch als Repertoire für die konfirmative Analyse eingesetzt werden - nicht nur für die Suche und Identifizierung von Mustern in den Daten, sondern für die Identifizierung von Mustern beim Abgleich.

3.1 Separation von Mustererkennung und Musterbeschreibung

Im letzten Kapitel wurden die Anforderungen an Visualisierungstechniken für die Präsentation bekannter Fakten und Zusammenhänge unterschieden von den Anforderungen an Visualisierungstechniken für die explorative Datenanalyse. Im Kern beruht der Unterschied darauf, dass eine Präsentation fokussiert werden kann auf die Inhalte, die damit kommuniziert werden sollen, während die Anforderung in der explorativen Datenanalyse mindestens anfangs darin besteht, einen möglichst umfassenden Überblick über alle Daten, Strukturen und Abhängigkeiten zu erhalten. Die Komplexität der Information, die dargestellt werden soll, ist in der explorativen Datenanalyse daher a-priori nicht bekannt. Dennoch kann jede Visualisierung stets nur einen bestimmten Anteil dieser Komplexität wiedergeben. Dabei ist sie unter anderem folgenden Einschränkungen unterworfen:

1. *Konstruktion der visuellen Abbildung.* Jede Visualisierungstechnik ist im Design eingeschränkt durch die zwei Dimensionen und die Größe des Bildraums. Ebenso schränken die Abhängigkeiten bei der Wahrnehmung verschiedener visueller Attribute [War04b, Ber83] die Wahl dieser Attribute erheblich ein. Hinzu kommt, dass alle Datenattribute in der explorativen Datenanalyse potentiell gleichberechtigt sind. Eine Auswahl von Datenattributen auf der Basis von nicht geprüften Annahmen wäre ein methodischer Fehler. Die Anzahl der Datenattribute, die gleichzeitig dargestellt und so in Beziehung gesetzt werden können, ist dementsprechend immer limitiert.

2. *Interpretation eines Musters.* Konkret geht es dabei um die Verbalisierung eines Musters im Sinne der Modelle von Lam und Ware. Die Verbalisierung ist kein präattentiver, sondern ein kognitiver Prozess, der daher sequentiell abläuft und das Arbeitsgedächtnis in Anspruch nimmt. Da dessen Kapazität eingeschränkt ist, ist auch die Komplexität der „lesbaren“ Informationen beschränkt - unter der Vorraussetzung, dass keine weiteren Hilfsmittel zur Verfügung stehen.

Die erste Einschränkung bezieht sich auf die Fähigkeit des Menschen, präattentiv Muster zu erkennen. Mit vielen, wenn nicht allen Visualisierungstechniken wird der Anspruch erhoben, dass die *Suche* nach Mustern oder interessanten Merkmalen eines Datensatzes als präattentiver Prozess unterstützt wird, der mit möglichst wenigen kognitiven Ressourcen auskommt. Hingegen erfordert die zweite Einschränkung, die Interpretation eines Musters im Anwendungskontext, dass entweder die Bedeutung der verwendeten visuellen Attribute vom Nutzer bereits gelernt wurde, oder dass die Bedeutung der Attribute durch bekannte, jedoch „arbiträre“ Symbole vermittelt wird.
Nach den Modellen der Informationsvisualisierung werden Mustererkennung und Musterbeschreibung mit den Techniken gleichermaßen unterstützt. Die Voraussetzung dafür, dass die Bedeutung der visuellen Attribute gelernt werden kann, ist aber, dass diese sich nicht ändert. Ein Beispiel dafür sind die „kanonischen“ Symbole und Farbgebungen in der Ikonographie vieler Landkarten (siehe u.a. [Ber83]): Karten desselben Typs (bspw. topographische Karten oder Autokarten) sind leichter zu nutzen, wenn die Interpretation im Sinne einer Übersetzung von sensorischen in arbiträre Symbole nicht für die jede Problemstellung neu gemacht werden

muss. Wenn eine Technik oder ein Visualisierungssystem dagegen die freie Zuordnung von Datenattributen auf visuelle Attribute unterstützen muss, müssen die für die Interpretation notwendigen Informationen symbolisch dargestellt werden.
Selbst wenn die Symbole ebenfalls auf dem Bildschirm angezeigt würden (etwa als Legende oder Beschriftung), müssen sie dennoch gelesen werden. Die Interpretation einzelner Symbole und die sich daran anschließende Interpretation des Musters ist ein sequentieller, kognitiver Prozess, der von der Suche nach einem Muster durch das Modell von Ware klar abgegrenzt werden kann. Die Beschreibung eines wahrgenommenen Musters nicht unbedingt eine Voraussetzung für eine Aktion oder Entscheidung durch den Menschen. Experten in einem Gebiet können auf eine Wahrnehmung eines Musters reagieren, ohne dass die Information dazu verbalisiert werden müsste. Allerdings funktioniert der Prozess der Wissensgenerierung nicht ohne die Beschreibung eines Musters, da die Muster nur auf diese Weise im analytischen Diskurs kommuniziert und damit exponiert werden können.
Die Unterscheidung zwischen der Erkennung und der Interpretation visueller Muster durch den menschlichen Betrachter berührt ein wichtiges Problem der Informationsvisualisierung: Die mögliche Komplexität der dargestellten Muster wird nicht nur durch die Fähigkeiten des Menschen eingeschränkt, sie als solche wahrzunehmen, als vielmehr durch seine Fähigkeiten, sie in gleichem Maße direkt zu interpretieren: Die Visualisierung unterstützt nicht die Verbalisierung. In Abschnitt 2.4.4 wurde bereits angesprochen, dass die Möglichkeiten, sich zu allen Datenelementen symbolische Detailinformationen anzeigen zu lassen, das Problem nicht löst. Schließlich kann das, was ein Muster charakterisiert, nicht nur als Summe seiner Elemente beschrieben werden.
Die mögliche Strategie, eine Visualisierung gerade so einfach zu gestalten, dass sie schnell zu lernen ist und die darin sichtbaren Muster direkt interpretierbar sind, wird in diesem Konzept nicht weiter verfolgt. Sie erscheint unter der Voraussetzung sinnvoll, dass der Suchraum in der explorativen Analyse auf einen entsprechend kleinen Ausschnitt beschränkt werden kann. In allen anderen Fällen widerspräche sie jedoch der Forderung, gerade die Aufgaben durch den Menschen durchzuführen, die uns leicht fallen und in denen wir automatische Methoden übertreffen.
Der Erkennung von Mustern ist ein Beispiel dafür. Aus dieser Forderung ergibt sich die Fragestellung, wie eine Strategie aussehen kann, so dass die Mustererkennung durch die Visualisierung unterstützt, die Musterbeschreibung jedoch durch andere Methoden umgesetzt wird. Integriert man Data-Mining und Machine-Learning in diesen Prozess, steht ein weiteres Repertoire von Techniken zur Verfügung, das ebenfalls für die Suche und Beschreibung eingesetzt werden kann. Dies rechtfertigt für ein Modell für Visual Analytics eine Verfeinerung des Visualisierungsprozesses, so dass auch solche Verfahren untersucht werden können, bei denen die beiden visuelle und automatische Techniken in einer engeren Verzahnung miteinander kombiniert werden können.

3.1.1 Rolle von Mustern im Data-Mining und in der Informationsvisualisierung

Eine Gemeinsamkeit zwischen den Modellen des Knowledge-Discovery und der Informationsvisualisierung besteht darin, dass die ersten Schritte der Pipelines praktisch identisch ablaufen. Die Datentransformation im Modell für die Informationsvisualisierung deckt eine Reihe entsprechender Prozesse in der KDD-Pipeline ab: Datenselektion, Datenaufbereitung und Transformation. Jeder dieser Aufgaben kann in beiden Modellen sinnvoll eingeordnet werden, und jede Technik dafür ist mit beiden Technologien einsetzbar, selbst wenn sie nur innerhalb eines dieser Forschungsgebiete entwickelt wurde.
Der Grund dafür ist, dass bis zur Datentransformation in beiden Pipelines keine Einschränkung darüber gemacht wird, welche Form die Ergebnisse dieser Prozesse tatsächlich haben. Auf dieser konzeptionellen Ebene sind die Ergebnisse der vorbereitenden Schritte gleich. Die gewählte Form und Beschreibung eines aufbereiteten Datensatzes ist nicht typisch für ein Verfahren aus der Informationsvisualisierung oder des Data-Mining. Aus diesem Grund wird im Rahmen dieser Arbeit davon ausgegangen, dass diese ersten Schritte nicht spezifisch sind für die Informationsvisualisierung oder den KDD-Prozess.
Daraus ergibt sich die Frage, ob sich die typisierenden Unterschiede zwischen den beiden Modellen im Kern auf die folgenden Schritte in den Pipelines reduzieren lassen könnten. Dies wären die *visuellen Abbildungen* und die *View Transformation* in der Informationsvisualisierung bzw. das *Data-Mining* im KDD-Prozess. Könnten diese Teilschritte gegeneinander ausgetauscht werden, ohne die Informationsvisualisierung und den Knowledge-Discovery-Prozess im Kern zu kompromittieren?
Aus der Sicht der Informationsvisualisierung ist das nicht möglich. Ihrem Namen und ihrer Definition nach erfordert die Informationsvisualisierung eine *visuelle* Repräsentierung der Daten. Im Vergleich zum Data-Mining ist das eine fundamentale Einschränkung, die nicht zuletzt durch die Architektur der Ausgabegeräte bestimmt wird: Auf den üblichen Displays ist die Repräsentierung nach dem Rendering letztlich ein zweidimensionales, trivariates Rasterbild.
Umgekehrt enthält die Definition des KDD-Prozesses das Ziel der Suche und Verifikation von Mustern - ein allgemeiner definiertes Ziel, das auch für die Rolle der Informationsvisualisierung im Analyseprozess gilt. In dieser Lesart wäre die Informationsvisualisierung ein Spezialfall des KDD-Prozesses, in dem die Suche der Muster nicht durch automatische Verfahren durchgeführt wird. Allein ihrer Funktion nach könnte man Data-Mining Methoden im KDD-Prozess auch durch entsprechende Visualisierungstechniken ersetzen. Dennoch gibt es einen wesentlichen Unterschied zwischen diesen Technologien, der von den Modellen für die Datenverarbeitungspipelines nicht betont wird.
Der Unterschied betrifft die jeweilige Definition eines Musters in jeder der Technologien. In beiden Fällen sind Muster nicht-beliebige Teilmengen einer Menge von Datenobjekten, die analysiert werden. Unterschiedlich sind aber die jeweiligen Kriterien für „Nicht-beliebigkeit“ in der Informationsvisualisierung und im Data-Mining. Die automatische Identifizierung oder Extraktion eines Musters aus einer Datenmenge erfordert die Spezifikation eines analytischen Modells in formaler Sprache. Diese Spezifikation enthält variable Symbole, d.h. Modellparameter, deren Werte erst mit Hilfe einer Heuristik bezogen auf die vorliegenden Daten optimiert werden müssen.

Viele automatische Verfahren erlauben die Konstruktion beliebig komplexer Modelle und damit die Beschreibung beliebig komplexer Muster (siehe Abschnitt 2.3.2). Durch ein Verfahren, das *alle* Teilmengen eines Datensatzes als Muster qualifizieren kann, würde die Auszeichnung als „Muster“ jedoch keine Aussagekraft haben. Um eine Abgrenzung zwischen Muster und Rauschen vorzunehmen, wird dementsprechend die Komplexität der Modelle den Gütekriterien nach beschränkt. Ein Modell, das von vornherein alle Muster erkennt, kann es in automatischen Verfahren nicht geben.
Im Knowledge-Discovery kann man aus der Definition von Mustern als „Ausdruck in einer Sprache, der eine Teilmenge der Daten beschreibt“ folgern, dass Identifikation und Beschreibung sich einander bedingen. Das bedeutet, dass die Wahl von Parametern und Modellen (bzw. Modellfamilien) die Mengen der „nicht-beliebigen“ Teilmengen determiniert. Dementsprechend werden andere Muster nicht gefunden werden. Im Gegensatz dazu funktioniert die Identifikation von visuellen Mustern im menschlichen Gehirn ohne die Spezifikation eines formalen Modells. Visuelle Muster genügen dieser Definition nicht, da sie wahrgenommen werden können, ohne dass sie dafür in natürlicher oder formaler Sprache beschrieben sein müssen.
In Abschnitt 2.4.4 wurde dargelegt, das die Identifikation und die Beschreibung eines visuellen Musters verschiedene Unterprozesse sind, für die unterschiedlich spezialisierte Hirnstrukturen zuständig sind. Der Wahrnehmung eines Musters entspricht der zweiten Stufe in Wares Modell, in dem alle Bildorte, die zum Muster gehören, zu einer Einheit zusammengefasst werden.

3.1.2 Erweiterung des Visual Analytics Prozessess

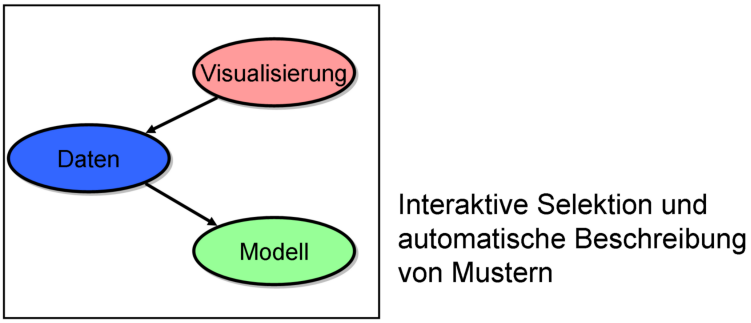


Abbildung 3.3: *In dieser Erweiterung wird eine Kopplungsvariante beschrieben, die darauf aufbaut, dass die Interaktion in der Visualisierung als Datenquelle für die automatische Analyse interpretiert werden kann.*

Betrachtet ein Nutzer eine Visualisierung und nimmt dabei ein Muster wahr, dann ist dies in den Modellen der Informationsvisualisierung und auch der Visual Analytics (siehe Abschnitte 2.4.3 und 2.5.1) nur mittelbarer Auslöser einer Interaktion. In diesen Modellen steht vor dem Eingriff des Nutzers mindestens seine Interpretation, d.h. eine abstrahierte Beschreibung des Musters. Diese Abstraktion ist einerseits notwendig, um das Muster im Anwendungskontext zu verstehen und bewerten zu können. Anderseits ist sie notwendig, um das wahrgenommene Muster mit den Interaktionsmöglichkeiten der Software in Beziehung zu setzen.

Lam [Lam08] beschreibt ein zyklisches Kostenmodell für die Nutzerinteraktion in Visualisierungssoftware. Dieses wiederum basiert auf Normans siebenstufigem, zyklischem Interaktionsmodell [Nor02]. Es setzt bei den Kosten für die Formulierung der Ziele an, identifiziert Kosten für den Eingriff des Nutzers für die Steuerung der Visualisierung und Kosten für die Interpretation und Evaluierung der „Reaktion“ des Systems. Ohne Änderung des Kostenmodells kann man ebenso beim Zyklus der Visualisierung beginnen, um die Kosten für die Umsetzung eines wahrgenommenen Bildes in eine Interaktion mit dem System zu beschreiben.
Lam diskutiert auch, inwiefern die Kosten für diese Umsetzung davon abhängig sind, wie vertraut die Nutzer im Umgang mit dem System sind. Dabei gibt es jedoch einen Unterschied zwischen den „Kosten für die Handlung“ und die „Kosten für die Bewertung“. Die Kosten für die Handlung können teilweise durch die Einhaltung von Konventionen beim Design von System und Steuerung effektiv verringert werden. Die Kosten für die Bewertung sind jedoch in der explorativen Datenanalyse nicht nur abhängig vom System, sondern auch abhängig von den Inhalten der Daten. Diese kann der Entwickler von vornherein natürlich nicht beeinflussen.
Die hier beschriebene Form der Kopplung zwischen automatischen und visuellen Methoden basiert darauf, dass die Interpretation und Evaluation eines Musters durch den Nutzer teilweise durch automatische Methoden ersetzt werden kann. Die Interaktion muss also auch dann noch funktionieren, wenn der Nutzer keine abstrahierte Beschreibung dessen besitzt, was als Muster auf dem Bildschirm wahrgenommen werden kann. Die Grundidee besteht darin, dass der Mensch-Maschine Dialog nicht nur auf einer abstrakten, symbolischen Ebene, sondern auch auf einer subsymbolischen Ebene unterstützt wird.
Überspitzt formuliert, beschreibt das Kostenmodell, dass der Nutzer dem System kommuniziert, was er *will* : Die Ausführung folgt zyklisch stets der Evaluation und Zielformulierung. Unter bestimmten Voraussetzungen ist stattdessen auch denkbar, dass der Anwender dem System nur kommuniziert, was er *sieht* (siehe Abbildung 3.4). Diese Anforderung beschränkt den Ort der Interaktion auf die Visualisierung selbst - und zwar genau jene, in der das Muster wahrgenommen wird. Eine Interaktion an jedem anderen Ort setzt voraus, dass der Anwender eine Korrespondenz zwischen dem Ort der Wahrnehmung und dem Ort der Steuerung herstellen kann.
Hier muss betont werden, dass aus dieser Einschränkung folgt, dass hier nicht alle denkbaren Kopplungen von visuell-interaktiven zu automatischen Verfahren betrachtet werden. Von einer Kopplung kann man beispielsweise auch dann sprechen, wenn ein Anwender in einer Visualisierung ein Muster erkennt, dieses interpretiert, evaluiert und auf der Grundlage der Evaluierung und der neu erworbenen Erkenntnisse die Parameter für die Steuerung des automatischen Verfahrens verändert. Abhängig davon, ob der Anwender die Korrespondenz zwischen Wahrnehmung und Steuerung durch eine Interpretation und Bewertung selbst herstellen muss, kann man von zwei verschiedenen Arten der Kopplung sprechen.
Eine Kopplungsvariante, bei denen das Korrespondenzproblem nicht allein durch den Menschen gelöst werden muß, folgt zwei Forderungen:

- Die Lösung des Korrespondenzproblems zwischen der Erkennung eines Musters und den darauf folgenden Schritten erforderte eine Arbeitsteilung zwischen Mensch und Maschine, die einseitig zu Lasten des Menschen geht. Die Arbeitsteilung beanspruch-

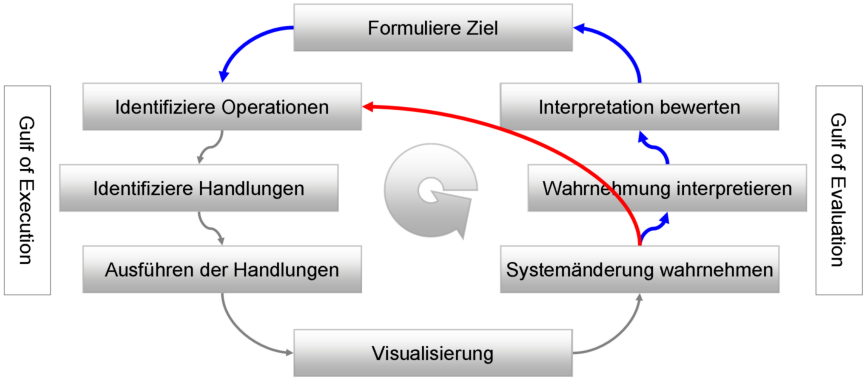


Abbildung 3.4: *Die Entkopplung zwischen Musterwahrnehmung und Musterbeschreibung beruht auf der Möglichkeit, den Interaktionszyklus abzukürzen (rot) und die Aktion des Nutzers unmittelbar an die Wahrnehmung zu koppeln. Dies bedeutet jedoch nicht, dass der Nutzer grundsätzlich keine Interpretation oder Evaluation vornimmt (blau). Die Voraussetzung dafür, dass Wahrnehmung und Handlung direkt gekoppelt werden können, ist, dass der Nutzer die Visualisierung kennt. Beim Lernen der Visualisierung muss vom Anwender ein Teil der Interpretationsleistung erbracht werden, die spezifisch ist für das Design, jedoch nicht spezifisch für ein bestimmtes Muster.*

te auch solche Fähigkeiten des Menschen, die dieser lernen muss bzw. nur mit hohem kognitiven Aufwand beherrscht. Hier wird stattdessen eine konservative Strategie verfolgt: Die Arbeit wird a-priori so aufgeteilt, dass dem Menschen nur solche elementaren Operationen zufallen, die er intuitiv beherrscht - in wesentlichen handelt es sich um das Erkennen und Vergleichen von Mustern. Auf dieser Grundlage wird untersucht, welche höheren Operationen allein durch die *technische* Kopplung zwischen visuell-interaktiven und automatischen Verfahren umsetzbar sind. Auf die kognitiven Fähigkeiten des Menschen soll dann und erst dann zurückgegriffen werden, wenn die Ausdrucksstärke möglicher Kopplungen nachweislich an die Grenzen stößt.

- Der zweite Grund betrifft den Anspruch von Visual Analytics als eigenständiges Forschungsgebiet, in dem nicht allein methodische, sondern auch neue *technische* Lösungen für die Probleme der Datenanalyse gefunden werden sollen. Man darf kritisch hinterfragen, ob eine nicht-technische Kopplung mit dem Nutzer als Mediator das Profil eines neuen technischen Forschungsgebietes begründet. Schließlich agiert der Mensch mindestens schon seit dem Beginn der elektronischen Datenverarbeitung in der Datenanalyse als Mediator zwischen unterschiedlichen Technologien und Werkzeugen. Der Anspruch, dass der Nutzer von dieser Vermittlungsaufgabe wenigstens teilweise entlastet wird, begründet hingegen tatsächlich das Profil eines neuen Forschungsgebiets. Die technische Kopplung zwischen Techniken der verschiedenen Forschungsgebiete muss letzten Endes einen Nutzwert erreichen, der größer ist als der Nutzwert der einzelnen Techniken (siehe Abbildung 3.5).

Inbesondere bei der Interpretation von Mustern soll der Anwender durch die automatische Modellierung entlastet werden. Die Kommunikation zwischen Mensch und Maschine auf subsymbolischer Ebene ist nur Mittel zum Zweck: Ziel ist nach wie vor eine formale Beschreibung der in den Daten identifizierten Muster unter der Vorrausetzung, dass die Aufgaben zwi-

schen Mensch und Maschine ihren Stärken entsprechend verteilt werden. Diese Erweiterung beschreibt einen mehrstufigen Prozess, der eine Visualisierungspipeline und ein Data-Mining Verfahren seriell miteinander verknüpft. Im einzelnen läuft diese Pipeline so ab:

1. Visualisierung
2. Musterwahrnehmung
3. Direkte Selektion (als subsymbolische Interaktion)
4. Musterbeschreibung (durch ein Data-Mining Verfahren)
5. Präsentation des Modells als Musterbeschreibung

Im folgenden werden die Schritte dieser Pipeline genauer beschrieben. Im Detail wird dabei besonders auf die Möglichkeiten der direkten Selektion für die subsymbolische Interaktion und die Anwendung von Data-Mining Verfahren eingegangen. Nur für sich betrachtet kann man diesen Prozess im Sinne der in Abschnitt 2.5.1 vorgestellten Taxonomie als „vorangestellte Visualisierung“ einordnen (siehe die Arbeit von Schneidewind [Sch07, Seite 39]). Hinsichtlich mehrerer Aspekte stellt dieser Teil des Konzepts dennoch eine Erweiterung dar:

- Durch die Unterscheidung symbolischer und subsymbolischer Interaktion, wobei Letztere Grundlage für die Entlastung des Nutzers von der Interpretation der Daten ist.
- Im folgenden Abschnitt werden verschiedene Möglichkeiten für die subsymbolische Interaktion und deren Verbindung mit Data-Mining Strategien erfasst. Ziel ist auch eine systematische Bestandsaufnahme der derzeit eingesetzten und auch der möglichen Kopplungsvarianten.
- Die Kategorisierung möglicher Kopplungen ist nicht allein deskriptiv zu verstehen. Die Arbeitsaufteilung zwischen Mustererkennung und Musterbeschreibung begründet ein Potential sowohl für die Spezialisierung von Visualisierungstechniken auf die Erkennung von Mustern, als auch für die Spezialisierung von Data-Mining Techniken auf die Modellierung von Mustern.

Der letzte Aspekt ist begründet auf zwei Beobachtungen: Die erste Beobachtung besteht darin, dass das Arbeitsgedächtnis die Komplexität der Zusammenhänge beschränkt [And96], die der Mensch zur gleichen Zeit verarbeiten und beschreiben kann. Unter diesen Umständen ergibt es jedoch keinen Sinn, Visualisierungen zu nutzen, die komplexere Zusammenhänge darstellen könnten. In der Tat setzen die (auch von Visualisierungsexperten) am häufigsten genutzten Visualisierungstechniken selten mehr als drei Attribute in Bezug[1].
Durch die Entlastung des Nutzers von der Musterbeschreibung kann diese Beschränkung aufgehoben werden. Im größeren Kontext stellt sich die Frage, warum die Visualisierungstechniken, die von Laien im Alltag genutzt werden, gerade nicht die sind, die im Fachgebiet

[1] Wie bereits beschrieben, bedeutet das nicht, das nur drei Attribute dargestellt werden (siehe Abschnitt 2.4.5).

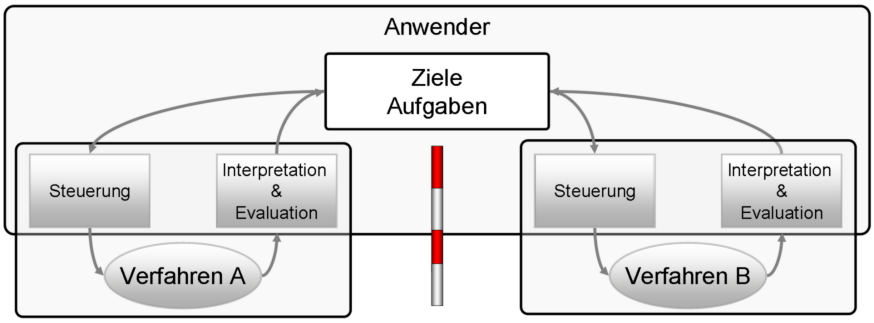


Abbildung 3.5: *Die Modelle von Norman und Lam für die Interaktion zwischen Mensch und technischem System beschreiben implizit auch eine Kopplungsmöglichkeit zwischen zwei verschiedenen Techniken oder Technologien. Durch Interpretation und Evaluierung werden die Ergebnisse einzelner Verfahren in den Kontext der Aufgabe des Anwenders übertragen. Jedoch handelt es sich dabei um eine methodische Kopplung, da die gekoppelten Techniken sich nicht direkt beeinflussen. Diese Art der methodischen Kopplung findet sich bei vielen Werkzeugen des Menschen. Stattdessen wird hier eine technische, direkte Kopplung zwischen den beiden Verfahren vorgeschlagen, um den Menschen von den Interpretationsaufgaben zu entlasten.*

als „State-of-the-Art“ gelten [vW05]. Die Darstellung komplexer Zusammenhänge mag das Design komplexerer Visualisierungen bedingen. Die Investition darin, eine neue Visualisierung zu erlernen, muss sich jedoch durch bessere, präzisere, sicherere Informationen amortisieren. Da die Komplexität der interpretierbaren Muster immer physiologisch beschränkt bleibt, würde sich diese Investition nicht lohnen - es sei denn, dass der Mensch die Interpretationsleistung nicht selbst erbringen muss.
Die zweite Beobachtung bezieht sich auf Data-Mining-Verfahren. Insbesondere Verfahren für das Clustering oder für die Erkennung von Ausreißern basieren auf der unüberwachten Erkennung von Strukturen und Mustern innerhalb des Datensatzes. Einerseits ist die Qualität der Ergebnisse stark abhängig vom zugrundeliegenden Modell, und andererseits kann diese - im Gegensatz zu überwachten Lernverfahren - nicht anhand von Referenzdaten überprüft werden. Die Robustheit von Data-Mining-Verfahren wird durch die vorgegebene Separation von Muster und Rauschen erheblich erhöht, da durch die Vorgaben der Suchraum für die Heuristiken eingeschränkt werden kann.
Die Tatsache, dass eine Visualisierungstechnik nur für die Erkennung von Mustern verwendet werden könnte, bedeutet nicht, dass existierende Techniken verändert werden müssten. Diesbezüglich soll hier betont werden, dass es sich um eine Erweiterung handelt, die existierende Modelle für die Interaktion und Kopplung ergänzt. Insofern wird auch keine Annahme über spezifische Eigenschaften der verwendbaren Visualisierungstechniken gemacht - abgesehen von der (nicht neuen) Voraussetzung, dass sie die Erkennung von Mustern effektiv unterstützt.
Die Pipeline unterscheidet sich am deutlichsten hinsichtlich der Kopplung von Interaktionstechniken mit Data-Mining-Verfahren. Im folgenden Abschnitt wird dieses Repertoire vorgestellt und beschrieben, auf welche Weise Mustererkennung und Musterbeschreibung über eine subsymbolische Interaktion miteinander gekoppelt werden können.

3.1.3 Interaktion als Datenquelle

Die hier vorgestellten Interaktionsmethoden sind nicht neu - vielmehr gehören sie zum Standardrepertoire der Graphischen Nutzerschnittstelle. Direkte Selektion und Direkte Manipulation (siehe Abschnitt 2.4.3.1) sind Interaktionsmethoden, die subsymbolische Interaktion erlauben.
Die Entwicklung neuer Interaktionsmethoden ist hier nicht das Ziel. Es soll stattdessen untersucht werden, welche Informationen der Nutzer direkt innerhalb einer Visualisierung beschreiben kann, und wie diese Informationen als Eingabe für automatische Verfahren genutzt werden können. Diese Untersuchung erfolgt in drei Abschnitten.
Die erste Stufe ist die *Charakterisierung der subsymbolischen Interaktionstechniken* und die Bestimmung der Art Informationen, die unmittelbar von Nutzer an die Maschine übergeben werden können. Eine Einschränkung ergibt sich insbesondere dadurch, dass die Interaktion innerhalb der Visualisierung stattfinden muss.
Die zweite Stufe ist die *Bestimmung des Urbilds der visuellen Abbildung*. Sie entspricht der Transformation der durch die Interaktion bestimmten Daten innerhalb des Modells, dass die visualisierten Daten beschreibt. Es wäre technisch durchaus denkbar, dass die Analyse auch direkt auf den untransformierten Bilddaten ausgeführt werden kann. Die automatische Analyse der Bilder - d.h. durch Bilderkennungsverfahren - würde den Menschen gerade bei der Mustererkennung aus dem Prozess ausschließen, wo dieser eine seiner größten Vorteile gegenüber automatischen Verfahren ausspielen könnte. Diese Möglichkeit wird daher für dieses Konzept nicht in Betracht gezogen.
Welche Informationen nach dieser Transformation vorliegen, hängt nicht mehr nur ab von den Interaktionsmethoden, sondern auch von der Visualisierungstechnik. Eine Charakterisierung der potentiell verfügbaren Informationen ist notwendig, um zu bestimmen, wie die Interaktion als Datenquelle für die automatischen Verfahren charakterisiert werden kann.
Die dritte Stufe ist die Anwendung automatischer Verfahren. In diesem Konzept soll jedoch nicht die Kopplung auf der Ebene einzelner Techniken untersucht werden - dies würde den Rahmen dieser Arbeit sprengen. Stattdessen wird für die in Abschnitt 2.3.2 vorgestellten Einzelziele der Analyse (d.h. Klassifikation, Clustering, etc.) untersucht, welche der aus den vorherigen Schritten verfügbaren Daten für eine Kopplung in Frage kommen.
Die Untersuchung soll eine systematische Betrachtung der Möglichkeiten für die Trennung zwischen Mustererkennung und Musterbeschreibung. Sie schließt daher bereits existierende Techniken, die dieses Konzept auf einer der drei Stufen umsetzen, ausdrücklich mit ein.

3.1.3.1 Direkte Selektion als subsymbolische Interaktion

Subsymbolische Interaktion ist deshalb „subsymbolisch“, weil sie nur voraussetzt, dass ein Muster genau jenen Teil des Bildes repräsentiert, den es einnimmt. Dies ist eine denkbar schwache Voraussetzung. Ein visuelles Muster ist in erster Linie ein *räumlich* begrenzter, nicht notwendig zusammenhängender, Bereich des Sichtfeldes bzw. des Bilds. In der Wahrnehmung kann dieser Bereich vom Rest des Bilds abgegrenzt werden.
Dabei kann das charakteristische Attribut des Musters potentiell jedes beliebige visuelle Attribut sein. Für die räumliche Abgrenzung ist eine bewusste Wahrnehmung des charak-

teristischen Attributes oder gar die Interpretation nicht notwendig[2]. Durch subsymbolische Interaktion wird die räumliche Abgrenzung auf dem Bildschirm reproduziert, um visuelle Objekte oder Merkmale zusammenzufassen oder zu trennen.
Dies grenzt das Repertoire einsetzbarer Techniken stark ein. Bis auf Ausnahmen, die im folgenden erklärt werden, kann der Anwender lediglich durch direkte Selektion den oder die Orte im Bild markieren, an denen er das visuelle Muster wahrnimmt. Jede Interaktion an *irgendeinem* anderen Ort erforderte die Lösung eines Korrespondenzproblems, was wiederum eine Interpretationsleistung erfordern würde.
Um jedoch die Interpretationsleistung zu charakterisieren, die ein Verfahren für die automatische Analyse überhaupt übernehmen kann, ist eine feinere Unterscheidung dessen notwendig, was „Interpretation einer Visualisierung“ bedeutet.
Die Interpretationsleistung, die der Mensch beim Betrachten einer Visualisierung zu erbringen hat, kann in mehrere Teile zerlegt werden. Im Allgemeinen müssen diese Leistungen jedoch nicht alle gleichzeitig erbracht werden, und bedeuten daher aus diesem Grund nicht immer den gleichen vollen Aufwand. Nach Lams Modell erfordert eine Visualisierung nur dann einen Aufwand für die Interpretation, wenn sich diese ändert. Nur selten ändern sich alle Komponenten einer Visualisierung immer zur gleichen Zeit. Vielmehr lassen sich die Kosten für die Interpretation einer Visualisierung staffeln (siehe Abbildung 3.6). Im Einzelnen sind dies:

1. Interpretation der Visualisierung, d.h. aller visuellen Metaphern, die unabhängig sind von durch die Interaktion veränderlichen Parametern.

2. Interpretation aktuellen Konfiguration einer Visualisierung, d.h. der Effekte der Parameter, die durch die Interaktion geändert werden können.

3. Interpretation der Muster selbst.

Die Interpretationsleistung für ein Muster baut auf den Vorangehenden auf. Beispiele für die erste Stufe sind die Darstellung eines Koordinatensystems für eine *Scatterplot*-Visualisierung oder die *Node-Link*-Metapher für die Darstellung eines Graphen. Unter der Voraussetzung, dass die Visualisierung eindeutig über ihre Metaphern erkannt werden kann, müssen diese Komponenten nicht jedesmal neu interpretiert werden.
Allerdings dienen die Metaphern auf dieser Stufe nicht nur der Charakterisierung der Visualisierung. Im Falle des Koordinatensystems definieren sie beispielsweise zusätzlich eine nicht-triviale Korrespondenz zwischen räumlich getrennten Punkten im Bild: In jedem Scatterplot repräsentiert die Höhe eines Punktes und die entsprechende Höhe auf der Ordinate denselben Wert. Eine Korrespondenz, die bereits gelernt und internalisiert wurde, verursacht bei der Interpretation von Mustern keine oder nur noch vernachlässigbare Kosten. In einem solchen Fall kann die Korrespondenz beim Design der Interaktion genutzt werden.
Während die Achsen des Koordinatensystems ein Beispiel für eine Metapher für die Interpretation der Visualisierung sind, erfordern Skala, Einheiten oder Werte eine Interpretation auf

[2]Die Fähigkeit Muster zu erkennen, ohne dass das es charakterisierende Attribut erkennbar wäre, wurde auf die Spitze getrieben mit den Bildern der Buchreihe "Das magische Auge", in denen einfache stereoskopische Bilder in komplexen Texturen verborgen wurden.

der zweiten Stufe. Hierbei handelt es sich im die Interpretation der Effekte einer Interaktion im Sinne von Lams Modell. Diese Größen gehören nicht zu den charakterisierenden Komponenten eines Scatterplots, sondern sind abhängig von den Parametern der Visualisierung. Beim Scatterplot handelt es sich also insbesondere um die Belegung der beiden Achsen und der anderen visuellen Attribute.
Die Interpretation der wahrgenommenen Muster ist die letzte Interpretationsleistung vor der Evaluation. Da sich die Muster immer ändern, sobald sich die Datenquelle geändert oder eine Interaktion stattgefunden hat, verursacht diese Stufe die höchsten Kosten. Diese Stufe soll von der zweiten Stufe deshalb unterschieden werden, weil eine Änderung der Datenmuster nicht notwendigerweise ein unmittelbarer Effekt der Interaktion sein muss. Nutzt man Visualisierungstechniken beispielsweise beim Monitoring, sollte die Konfiguration deshalb seltener geändert werden, um einen einheitlichen Referenzrahmen zu schaffen, in dem Änderungen in den Daten leichter wahrnehmbar sind.
Insgesamt bauen alle einzelnen Interpretationsleistungen aufeinander auf. Im hier vorgestellten Konzept wird die Korrespondenz über alle drei Stufen hinweg automatisch hergestellt - jeweils auf der Basis der Visualisierungstechnik und ihrer aktuellen Konfiguration. Dass der Anwender die Visualisierung kennenlernt, und damit eine Interpretationsleistung für die ersten beiden Stufen bereits erbracht hat, ist jedoch Voraussetzung für die Ausführung der Interaktion. Diese Interpretationsleistung ist jedoch unabhängig von einem bestimmten Muster und unabhängig von den meisten Parametern der Visualisierung. Wie dieser Lernprozess für den Anwender insgesamt erleichtert werden kann, wäre Gegenstand weiterführender Forschung. Die Selektion von Punkten im Bild ist die direkteste Form der Interaktion mit der Visualisierung. Dabei kann man folgende Varianten unterscheiden:

- Direkte Selektion und Manipulation von visuellen Abbildern von Datenobjekten
- Markierung und Verschiebung von Punkten im visuellen Raum (d.h. Punkten oder Punktmengen, die nicht mit Datenobjekten assoziiert sind.)

Visuelle Abbilder von Datenobjekten liegen zwar auf Punkten des visuellen Raums, sie werden aber deshalb unterschieden, weil ihre Urbilder bezüglich der visuellen Abbildung im allgemeinen unterschiedliche Informationen beschreiben: Das Urbild eines visuellen Abbilds ist natürlich das Datenobjekt selbst. Das Urbild eines beliebigen Punktes im visuellen Raum kann abhängig von der visuellen Abbildung ein Punkt im Datenraum sein, eine beliebig geartete Punktmenge oder auch einen Wertebereich des Datenraums.
Für die räumliche Abgrenzung der Muster eignen sich noch weitere graphische Primitive wie Strecken und Geraden, Linienzüge oder Regionen. Eine Grenze bei der Definition dieser Primitive für eine Technik wird nicht durch die Erkennbarkeit der Muster bestimmt, sondern durch die Notwendigkeit, die Steuerung für die Definition dieser Primitive zu lernen. Streng genommen bezieht sich die Unterscheidung „symbolisch vs. subsymbolisch“ hier nur auf die Ebene des Kostenmodells, in dem Interpretation und Evaluation durchgeführt werden (*„gulf of evaluation“* [Lam08]). Die Ausführungsebene (*„gulf of execution“*) ist davon nicht betroffen und liegt auch nicht im Fokus dieses Konzepts. Ob eine entsprechende Charakterisierung auf der Ausführungsebene überhaupt möglich ist oder ob es eine Entsprechung gibt, wäre Gegenstand weiterführender Forschung.

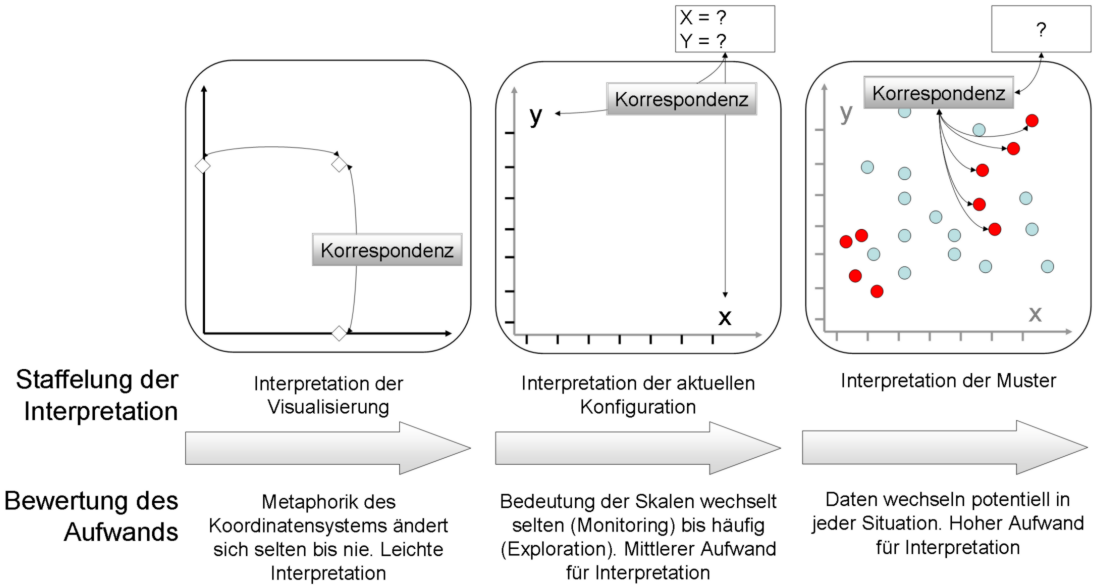


Abbildung 3.6: *Die Leistung, die für die Interpretation des wahrgenommenen Bildes erbracht werden muss, läßt sich staffeln. Sie muss nur jeweils für den Teil der Visualisierung erbracht werden, der noch nicht internalisiert wurde, bzw. für jenen Teil, der sich am häufigsten ändert. Wird beispielsweise die Visualisierung vom Betrachter als Streudiagramm erkannt, folgt daraus unmittelbar die Korrespondenz zwischen Punkten die untereinander und Punkten die nebeneinander liegen (links). Ist diese Korrespondenz bekannt und internalisiert, kann diese ohne Aufwand abgerufen werden - unabhängig von der verwendeten Skala und den Daten. Sobald sich die Skalen verändern, ändert sich auch die Bedeutung absoluter Positionen der Visualisierung, d.h. die Korrespondenz zwischen visuellen und symbolischen Artefakten (mittleres Bild). In bestimmten Anwendungen, etwa bei Monitoringaufgaben kann auch diese Korrespondenz internalisiert werden, indem die Skalen nur selten gewechselt werden. Die Erkennung der Muster selbst (rechts), bedeutet deshalb innerhalb der explorativen Analyse den höchsten Aufwand, weil sie Artefakte der Daten sind, die am häufigsten wechseln. Die Interpretationsleistung daher auch im häufigsten neu erbracht werden.*

Es wird daher im übrigen keine Annahme darüber gemacht, über welche Interaktionsmetapher oder welches Eingabegerät die Interaktion durchgeführt wird; in erster Linie deshalb, weil sonst die Unterscheidung zwischen den Primitiven nicht eindeutig wäre. Beispielsweise ist eine Menge von Datenobjekten, die als Punkte dargestellt wird, durch Einkreisen mit einem Linienzug auswählbar. Für Datenobjekte, die als Linienzüge dargestellt sind (etwa bei Holten [HvW08]), beschreiben Konyha et al. [KMG+06] *Line-Brushes* - eine Technik, in der die Datenobjekte durch Überstreichen mit dem Zeigegerät selektiert werden. Die Unterscheidung zwischen den Primitiven bezieht sich ausschließlich auf die *Urbilder* der visuellen Abbildung, denn dies sind die für das automatische Verfahren relevanten Informationen.

Die räumliche Abgrenzung der wahrgenommenen Muster innerhalb der Visualisierung sind die einzigen Informationen, die der Nutzer ohne vorherige Interpretation für die Musterbeschreibung liefern kann. Dies heißt nicht, dass dies die einzigen Informationen sind, die weiterverarbeitet werden dürfen. Alle Parameter, die die Transformationen der Visualisierung steuern, können in der Folge verwertet werden. Insbesondere definieren diese Parameter auch die Umkehrabbildung vom Bildraum in den Merkmalsraum der dargestellten Daten, die der erste Schritt für die automatische Interpretation der Muster ist. Dabei ist es unerheb-

lich, ob diese Parameter automatisch bestimmt wurden, oder ob sie vom Nutzer bewusst mit einer bestimmten Absicht gewählt oder getestet werden. Der Fokus auf die subsymbolische Interaktion schränkt das Repertoire „erlaubter“ Interaktionsmetaphern nur insofern ein, da der Anwender kein für ihn neues Korrespondenzproblem lösen soll.
Diese Einschränkung ist im folgenden jedoch deshalb notwendig, um im Rahmen dieses Konzepts zu bestimmen, wie die Trennung zwischen Mustererkennung und Musterbeschreibung umgesetzt werden kann, und welche weiteren Konsequenzen sich in den folgenden Verarbeitungsschritten aus dieser Einschränkung ergeben.

3.1.3.2 Urbilder der visuellen Abbildung

Die Interaktion selbst operiert bei den gegebenen Einschränkungen ausschließlich innerhalb der Visualisierung. Der ersten Teil der Interpretationsleistung - die ja anstelle des Menschen durch die Maschine übernommen werden muss - muss daher innerhalb der Visualisierung stattfinden. Im Abschnitt 2.4.3.1 wurde bereits beschrieben, dass dadurch die Art der Informationen, die durch direkte Selektion beschrieben werden kann, von der Visualisierung abhängig ist.
Die Unterscheidung von intensionalen und extensionalen Urbildern (nach Derthick et al. [DKR07]) der visuellen Abbildung wird in diesem Konzept übernommen. Ein extensionales Urbild ist die Menge aller ausgewählten Datenobjekte in der Form einer Auflistung ihrer Identifikatoren. Ein intensionales Urbild repräsentiert eine Teilmenge des Merkmalsraums. Im Sinne von Hands Definition (siehe Abschnitt 2.3.2) handelt es sich bei intensionalen Urbildern um Modelle. Die Transformation einer extensionalen Beschreibung in eine intensionale Beschreibung ist die Domäne des Data-Mining, was eine Kopplung mit entsprechenden Verfahren motiviert.
Die umgekehrte Transformation entspräche einer Suchanfrage etwa in einer Datenbank. Wenn bereits die Visualisierungtechnik eine hinreichend einfache und präzise intensionale Beschreibung (als Urbild eines Musters) liefert, ist eine Kopplung zwischen Visualisierung und Data-Mining Verfahren allein für die Beschreibung nicht zwingend notwendig.
Ein intensionales Urbild unterscheidet sich formal in keiner Weise von einem Modell im Sinne des Data-Mining. Alle Verfahren, die etwa eine direkte *Dynamic Query* umsetzen [AWS92], d.h. eine Suchanfrage, in der ein Intervall von Werten innerhalb der Visualisierung beschrieben werden kann, erzeugen ein entsprechendes Modell. Unabhängig davon, dass sie eigentlich für die Formulierung von Suchanfragen entwickelt wurden und werden, lassen sich diese Modelle natürlich auch als binäre Klassifikationsmodelle auffassen.
Dies mag einer der Gründe dafür sein, warum die Einordnung von Arbeiten zwischen den Forschungsgebieten Informationsvisualisierung und Visual Analytics nicht immer scharf getrennt werden kann. Eine Abgrenzung nach der Komplexität der Modelle (d.h. der Anzahl ihrer Parameter) wäre nur bedingt geeignet, da Ansätze aus der Informationsvisualisierung existieren, mit denen im Prinzip beliebig komplexe Modelle konstruiert werden können. Beispiele, in denen diese Modelle Queries definieren, sind zum Beispiel das *Compound-Brushing* von Chen [Che04], oder die *Timebox-Widgets* von Hochheiser und Shneiderman [HS04]. In beiden Fällen werden durch die Interaktion zusammengesetzte Queries definiert. North und Shneiderman erweitern in [NS00] diese Konstruktion auf mehrere verbundene Visualisierungstechniken.

Die Art der Modelle, die direkt als Urbildmenge eines gegebenen Repertoires an Interaktionsprimitiven beschrieben werden kann, hängt von der visuellen Abbildung ab. Vorgegeben ist, dass die Interaktion lediglich räumlich definierbare Primitive erzeugen kann. Dementsprechend definieren in allen Visualisierungstechniken genau die Datenattribute die Urbildmenge, die auf räumlich abgrenzbare visuelle Attribute abgebildet werden.
Gibt die Visualisierungstechnik beispielsweise zwei feste, orthogonale Achsen für je ein Datenattribut vor, ist die Urbildmenge eine Punktemenge, die durch diese Attribute bestimmt wird. Bei linearen Projektionstechniken kann die Urbildmenge dementsprechend mehr Attribute umfassen. Bei nicht-linearen Projektionen schließlich kann die Urbildmenge noch vielgestaltiger sein. Eine formale Grenze für die Gestaltbarkeit der Modelle gibt es in der Informationsvisualisierung ebenso wenig wie im Data-Mining. Zum Beispiel kombiniert die Visualisierungstechnik von Walter et al. [WOWR03] Techniken des *Multidimensional Scaling* mit hyperbolischen Projektionstechniken. Eine Umkehrabbildung für einen zusammenhängenden Bereich - selbst wenn er im Bildraum einfach definierbar ist - kann in diesem Fall eine beliebig komplexe Urbildmenge erzeugen.
Eine grundsätzliche Charakterisierung und Abgrenzung zwischen intensionalen Modellen aus der Informationsvisualisierung und Modellen aus dem Data-Mining ist für dieses Konzept nicht relevant. Wichtig ist dagegen, dass die Informationen, die durch die Interaktion definiert werden, potentiell kompatibel sind zu den Eingaben, die für die Steuerung der Data-Mining-Verfahren notwendig sind. Die Möglichkeiten für die Kopplung zwischen interaktiven und automatischen Verfahren potenzieren sich dadurch, dass durch die Visualisierung nicht vorgegeben ist wie die Informationen letztlich interpretiert werden.

3.1.4 Nutzung der Daten aus der Interaktion in automatischen Verfahren

Eine Visualisierungstechnik, ihre visuelle Abbildung und ihre direkten Interaktionsmethoden bestimmt das Repertoire von Informationen, dass für die Analyse zur Verfügung gestellt werden kann. Im Visual-Analytics-Prozess (siehe Abschnitt 2.5.1) dient die Analyse der Konstruktion von Modellen. Aus der generischen Beschreibung für ein Data-Mining-Verfahren (2.3.4) lassen sich drei mögliche Arten ableiten, wie diese Informationen durch ein automatisches Verfahren interpretiert werden können (siehe Abbildung 3.7).
Die erste Möglichkeit ist die Interpretation der Informationen als Eingabedaten für die Analyse. In diesem Fall wird durch die Interaktion wird ein neuer Datensatz erzeugt oder modifiziert, der genauso behandelt werden kann, wie jeder Datensatz aus dem gleichen Merkmalsraum. Im Sinne des Modells für den Visual-Analytics-Process wird ein neuer Datenfluss etabliert, der von der Visualisierung zurückführt auf eine neue Datenquelle (siehe Abbildung 3.3). Das automatische Analyseverfahren liefert eine formale Beschreibung des Musters. Da durch die Interaktion stets eine Teilmenge des ursprünglichen Datensatzes beschrieben wird, gibt es durch die Kopplung keine zusätzlichen technischen Einschränkungen. Jedes automatische Verfahren, das auf den ursprünglichen Daten operiert, kann auch für eine Teilmenge dieses Datensatzes eingesetzt werden.
Die zweite Möglichkeit ist die Interpretation der Eingabedaten als *Modellparameter*. In diesem Fall können die Modellparameter verändert werden. Dies bedeutet nicht zwangsläufig,

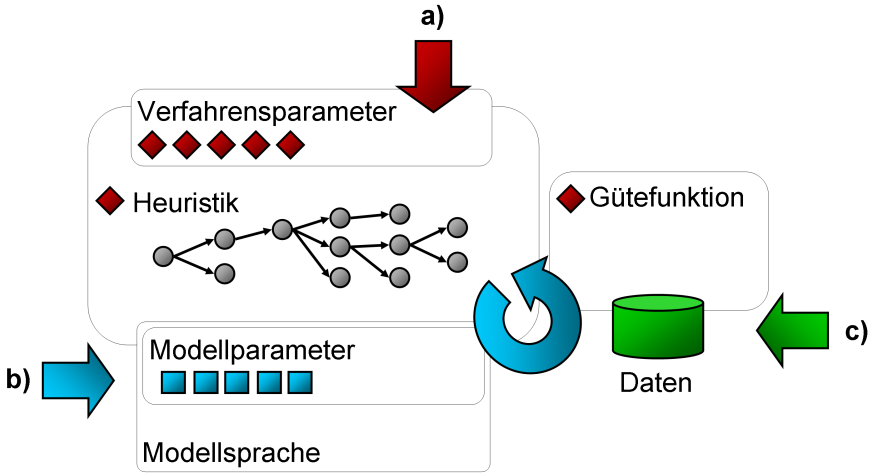


Abbildung 3.7: *Für die Kopplung von automatischen Techniken lassen sich drei verschiedene Ansatzpunkte unterscheiden: Die Verfahrensparameter (a), die Modellparameter (b) und die Daten (c). Auch Gütefunktion und Heuristik können als Verfahrensparameter betrachtet werden, da sie häufig in verschiedenen Varianten ausgetauscht werden können. Da die Gütefunktion unter anderem über die Daten definiert ist, hat jede interaktive Änderung der Daten auch einen Einfluss auf die Eigenschaften der Gütefunktion.*

dass ein automatischer Algorithmus entbehrlich wird. Erstens ist nicht ausgeschlossen, dass der Nutzer nur einen Teil der Parameter über die Interaktion definieren kann. Zweitens ist nicht auszuschließen, dass der Nutzer durch die Wahl der Modellparameter ein Modell konstruiert, das bezüglich der gegebenen Gütefunktion nicht optimal ist. Eine Kopplung ist sowohl dann möglich, wenn die Modellparameter direkt gesetzt werden, als auch dann, wenn die Eingabedaten für eine Initialisierung der Heuristik genutzt werden.
Die dritte Möglichkeit ist die Interpretation der Eingabedaten als *Verfahrensparameter*, die die Freiheitsgrade eines automatischen Verfahrens beschreiben, deren Werte durch den Entwickler des Verfahrens nicht a-priori festgelegt wurden.

Die Unterscheidung ist so nicht für jedes Verfahren eindeutig. Beispielsweise sind Samplebasierte Klassifikationstechniken, etwa die *Nearest-Neighbor-Klassifikation*, gerade dadurch charakterisiert, dass die Eingabedaten - die Samples - unverändert als Modellparameter interpretiert werden. Im Prinzip sind in diesem Verfahren Eingabedaten und Modellparameter nicht zu unterscheiden. Ein anderes Beispiel ist das *K-Means* Clusteringverfahren. Der Parameter K ist insofern ein Verfahrensparameter, weil dieser a-priori gegeben wird, und die Heuristik für die Suche der Cluster steuert. Da K die Anzahl der Cluster beschreibt, ist er gleichzeitig die wichtigste charakteristische Größe des Modells.
Diese Mehrdeutigkeiten werden im Rahmen dieser Typologie folgendermaßen aufgelöst: Wenn die durch die Interaktion definierbare Information nach dem Verfahren unverändert im analytischen Modell verwendet wird, dann gelte diese Information in jedem Fall als Modellparameter. Da die meisten Verfahren durch mehr als eine Klasse von Parametern gesteuert werden, kommen entsprechend häufig mehr als nur eine dieser Möglichkeiten in Betracht.
Die sich daraus ergebende Charakterisierung unterscheidet zwei für die Kopplung wichtige Merkmale von Eingabedaten für analytische Verfahren. Das erste Merkmal unterscheidet Modellparameter und Eingabedaten danach, ob eine (beliebige) Punktmenge eines Merk-

malsraums *direkt* oder *indirekt* für die Definition des Modells verwendet wird. Gemeinsam ist beiden, dass die Informationen durch subsymbolische Interaktion in der Visualisierung einfach definiert werden können. Sie unterscheiden sich aber darin, wie viel direkten Einfluss der Mensch auf die Definition des Modells nimmt. Dies ist deshalb bedeutsam, weil in den Verfahren, in denen der Anwender *jeden* Modellparameter direkt bestimmen kann, eine automatische Modellbeschreibung nicht stattfindet. Auch wenn entsprechende Ansätze eine große Flexibilität versprechen, kann man nicht von einer Kopplung der beiden Technologien sprechen.
Van Wijk [vW05] schreibt, dass weder Visualisierung noch Interaktion per se „gut“ sind. Interaktion sollte so eingeschränkt werden, dass Modelle nicht beliebig gestaltet werden können, weil durch die automatische Optimierung garantiert ist, dass die Konstruktion des Modells expliziten, nachvollziehbaren Kriterien - den Gütefunktionen - folgt. Das bedeutet auch, dass die Beschreibung eines Modells ein wiederholbares Experiment bleiben muss. Beim visuellen Feedback (siehe Abschnitt 3.2.3) manifestiert sich diese Anforderung in der Information, die der Nutzer durch die Interaktion über das Modell gewinnen kann. Bei dem Grad der Einflussnahme des Nutzers auf die Modellparameter gibt es verschiedene Abstufungen, abhängig von Grad der Indirektion, den ein Verfahren bei der Interpretation der Parameter umsetzt. Die Wahl von Verfahrensparametern hat immer einen indirekten Einfluss auf das entstehende Modell. Beispiele dafür sind Verfahrensparameter, die beim *Pruning* von Entscheidungsbäumen bestimmen, wie komplex der Baum werden darf. Das gleiche gilt für komplexere Informationen wie Distanz- oder Qualitätsmaße und Verteilungsfunktionen, die in praktisch allen Data-Mining oder statistischen Verfahren definiert werden müssen. Verfahrensparameter haben unter Umständen keinen oder einen nicht-trivialen Bezug zum Merkmalsraum in dem die Daten liegen.
Das bedeutet, dass die Kopplung zwischen automatischen und visuell-interaktiven Verfahren nicht für alle Parameter notwendigerweise eine Visualisierung der Daten nutzen muss. Wenn die Wahl eines Verfahrensparameters beispielsweise weniger abhängig von den a-priori unbekannten Strukturen der Daten ist, sondern vielmehr abhängig ist von den Zielvorgaben der Problemstellung ist, dann ist auch eine Visualisierung der Verfahrensparameter im Kontext dieser Zielvorgaben denkbar (wie zum Beispiel Receiver-Operating Characterisics (ROC)).

Wie die gegebenen Informationen verarbeitet werden, hängt natürlich vom Verfahren selbst ab, kann in vielen Fällen jedoch auch über die Problemklasse innerhalb der Analyse allgemeiner beschrieben werden. Dabei sollen folgende Problemklassen untersucht werden:

- Klassifikation & Regression
- Clustering
- Erkennung von Ausreißern
- Beschreibende Zusammenfassung

Diese Problemklassen korrespondieren direkt mit einer Klasse von Data-Mining oder statistischen Verfahren. Innerhalb des KDD-Prozess gibt es daneben jedoch vorbereitende, nichttriviale Arbeitsschritte, die das Ergebnis der automatischen Prozesse mittelbar beeinflussen.

Die Frage, wie diese Arbeitsschritte mit Visualisierungstechniken gekoppelt werden können, und wie diese Kopplung eine bessere Grundlage für die entsprechenden Entscheidungen bietet, stellt sich sich auch bei den folgenden Problemklassen

- Aggregation
- Attributauswahl
- Dimensionsreduktion
- Parameterauswahl
- Distanzmaße

Auch diese Aufgaben erfordern möglicherweise Informationen, die innerhalb der Analyse aus den Daten erst gewonnen werden müssen. Daher erscheint es auch hier sinnvoll, zu untersuchen, wie Visualisierung mit diesen Aufgaben integriert werden kann. Für jede Problemklasse wird allgemein beschrieben, welche Informationen die Verfahren für diese Problemklasse benötigen. Alle Daten oder Parameter stellen eigenständige Ansatzpunkte für eine Kopplung dar. Bei der folgenden Beschreibung werden besonders die Ansätze berücksichtigt, in denen Daten oder Parameter für ein automatisches Verfahren *durch eine Visualisierung* interaktiv definiert werden können.
Die anderen beiden Optionen für die Eingabe wären

- die automatische Bestimmung von Daten oder Parametern durch vorbereitende Verfahren (wie beispielsweise in parameterfreien Verfahren)
- die Eingabe durch den Nutzer über GUI-Elemente (nach Interpretation und Evaluation).

Bei dieser Untersuchung sollen auch bereits existierende Ansätze untersucht werden. Dabei sollen hier nur solche berücksichtigt werden, in denen die Informationsvisualisierung in erster Linie für die Wahrnehmung der Muster genutzt wird. Unberücksichtigt bleiben daher insbesondere Techniken der (reinen) Informationsvisualisierung, in denen der Anwender die Interpretationsleistung für die Modellierung vollständig selbst erbringen muss.

3.1.4.1 Klassifikation & Regression

Bei der Klassifikation handelt es sich um die Bestimmung einer Funktion $f : X \to Y$, wobei die Eingabedaten eine Menge von Tupeln $(x, y) \in (X \times Y) \subseteq M$ beschreiben. Y ist eine Menge von Klassen und M der Merkmalsraum, der alle Attribute eines Datensatzes umfasst und innerhalb dem das Klassifikationsmodell gelten soll.
Da es sich bei der Klassifikation um überwachte Lernverfahren handelt, ist die wichtigste Eingabe eine Menge von Trainingsdaten $S \subseteq X \times Y$, deren Klassifikation von vornherein bekannt ist. Durch die direkte Interaktion mit einem Display, das die Datensätze des Merkmalsraums darstellt, ist es möglich, eine binäre Klassifikation zu definieren. Im einfachsten

Fall werden durch die Selektion eines Musters $P \subset X$ manuell die Klassen $Y = \{P, X \backslash P\}$ erzeugt. Abhängig davon, welche Komplexität man bei der Bedienung des User-Interfaces zuließe, wäre es möglich, beliebig viele Klassen zu definieren. Andererseits kann man mehr Klassen auch durch die einzelne Bearbeitung und automatische Zusammenfassung mehrerer binärer Klassifikatoren erreichen. Durch eine solche Strategie kann man die Bedienkomplexität vorgeben, ohne grundsätzliche Einbußen an Funktionalität zu erleiden.

In vielen Verfahren werden die Klassifikatoren direkt interaktiv modelliert; die Interaktion innerhalb der Datenvisualisierung wird also direkt als Modellparameter interpretiert. An dieser Stelle sollen nur solche Ansätze betrachtet werden, in denen die Klassifikationsmodelle durch die Interaktion innerhalb klassifizierter Daten modelliert werden. Die andere Möglichkeit, beispielsweise einen Entscheidungsbaum innerhalb der Visualisierung seiner Hierarchie direkt zu verändern, ist eine Interaktion innerhalb des Modells und Gegenstand von Abschnitt 3.1.5.
Entscheidungsbäume eignen sich deshalb besonders gut für die interaktive Modellierung, da die Komplexität einer einzelnen Entscheidung auf wenige Attribute des Merkmalsraums reduziert werden kann. Der Preis dafür ist, dass bei der Modellierung von Entscheidungsbäumen zwei Probleme gelöst werden müssen. Das erste Problem ist die Wahl der Attribute, die für jeden Knoten relevant sind, das zweite Attribut ist die Modellierung des Klassifikators für einen Knoten. Bei der Modellierung eines Entscheidungsbaums kann für diese beiden Probleme unabhängig entschieden werden, ob sie jeweils automatisch oder interaktiv gelöst werden.
Zu den frühesten Ansätzen für die Konstruktion eines Entscheidungsbaums gehört die Arbeit von Ankerst et al. [AEEK99]. Diesen Ansatz zeichnet insbesondere aus, dass die Darstellung der Daten und die Darstellung des Modells eng miteinander verbunden wird. Ein Knoten des Baums stellt die Verteilung der einzelnen Klassen bezüglich des Wertebereichs für jeweils ein gewähltes Attribut dar. Die Attributwerte, an denen sich die Verteilung der Klassenhäufigkeiten wahrnehmbar ändert, sind Kandidaten für die Unterteilung der Wertebereiche.
Die rekursive Unterteilung des Entscheidungsbaums wird in diesem Ansatz manuell umgesetzt. Die Herstellung der Korrespondenz zwischen Merkmalsraum und Modellraum wird dabei effektiv unterstützt. Da alle Attribute des Merkmalsraums gleichzeitig dargestellt werden können, kann auch die Wahl des Attributes für einen Knoten interaktiv auf der Basis der Visualisierung erfolgen.
Ankerst et al. untersuchen in einer Folgearbeit [AEK00] auch mehrere Strategien für eine Kopplung bei der Konstruktion von Entscheidungsbäumen. Eine Strategie besteht darin, dass über ein Standardverfahren Vorschläge für potentielle Unterteilungen gemacht werden, die der Nutzer annehmen oder ablehnen kann. Eine zweite Strategie besteht darin, dass es von der Tiefe des Baums abhängt, ob ein Modellparameter automatisch oder interaktiv bestimmt wird. In einer Variante wird der Baum durch den Nutzer initialisiert, d.h. die ersten Knoten werden manuell definiert und anschließend wird der Baum automatisch weiterberechnet. In einer zweiten Variante wird es umgekehrt gemacht. Dies sind zwei Möglichkeiten, die zeigen, dass die Arbeitsteilung zwischen Mensch und Maschine auch über Modellparameter der gleichen Art funktionieren kann.
Ware et al. [WFH+01] entwickeln ein System für die Konstruktion eines binären Entscheidungsbaums. Auch in dieser Variante werden die Modellparameter direkt durch die Interaktion innerhalb der Visualisierung bestimmt. Da die Knoten in diesem Modell die Klassen in

zwei numerischen Dimensionen separieren, werden die Klassifikatoren innerhalb von *Scatterplots* modelliert. Das System beruht auf der iterativen Modellierung durch klassentrennende Polygone, die der Nutzer jeweils in den *Scatterplots* nachzeichnet. Parallel dazu wird der entstehende Entscheidungsbaum in einer zusätzlichen Ansicht dargestellt - in diesem Fall getrennt von den eigentlichen Daten. In diesem Fall gibt der Anwender durch die Interaktion keine Eingabedaten vor, sondern beschreibt die Modellparameter - die trennende Hyperebene - direkt. In beiden Verfahren wird der Entscheidungsbaum manuell aufgeteilt, wodurch das Problem der Überanpassung umgangen werden kann. Die Auswahl von Attributen wird mit einer ähnlichen Visualisierung unterstützt wie bei Ankerst.
Zwei weitere Beispiele für semi-automatische Techniken sind die Arbeiten von Teoh und Ma [TM03] und Liu et al. [LSG04]. Bei Teoh und Ma wird ein Entscheidungsbaum konstruiert, allerdings ist die Anzahl der Attribute, die für einen Knoten des Baums relevant sind, nicht beschränkt. Die Arbeit von Liu et al. geht besonders weit hinsichtlich der Möglichkeit, die Modelle direkt zu editieren. Es erlaubt die Beschreibung von Klassifikationsregeln auf der Basis von Vereinigungsmengen konvexer Polytope im hochdimensionalen Merkmalsraum (als Entscheidungsbaum hätte dieses Modell die Tiefe eins).
Dieses Modell erlaubt die Beschreibung fast beliebig komplexer Muster durch den Anwender; jedoch ist auch bei der manuellen Konstruktion durchaus die Gefahr einer Überanpassung gegeben. Nicht alle diese Verfahren sind unverändert für eine Kopplung mit automatischen Verfahren brauchbar. Wie im vorigen Abschnitt beschrieben, muss es für eine Kopplung möglich sein, dass mindestens ein Teil der Modellparameter über das automatische Verfahren bestimmt werden kann, so dass Freiheitsgrade für die Optimierung bleiben.

In den bisher genannten Verfahren werden die Muster direkt durch deren Grenzen im Merkmalsraum der Daten definiert. Die folgenden Verfahren unterschieden sich darin, dass die räumliche Abgrenzung eines Musters durch den Anwender implizit definiert wird. Der Nutzer klassifiziert dabei eine Menge von Trainingsdaten derart, dass diese einem wahrgenommenen Muster entspricht. Die Freiheitsgrade, die für die automatische Optimierung des Modells verbleiben, sind die gleichen, die den Algorithmen auch bei jedem anderen Trainingsdatensatz zur Verfügung stünden.
Die Interpretation der Eingabe als Trainingsdatensatz eignet sich deshalb gut für die Beschreibung visueller Muster, weil sie unabhängig ist von der Visualisierungstechnik und unabhängig vom verwendeten Klassifikationsverfahren. Verwendbar ist jedes Klassifikationsverfahren, das auf die Grundmenge X angewendet werden kann.
Im folgenden Kapitel (siehe Abschnitt 4 und [MK08b]) wird eine Form der Umsetzung dieses Konzepts vorgestellt, bei dem ein ausgewähltes Muster die Eingabe für einen Entscheidungsbaum darstellt. Der gelernte Klassifikator führt eine Separation zwischen dem Muster und dem Rest der dargestellten Daten durch. Der Klassifikator kann sowohl als prädiktives als auch als deskriptives Modell interpretiert werden. Die extensionale Darstellung des Musters - die selektierte Menge - wird in nicht trivialer Form in eine Beschreibung überführt.
Ein auf dieser Ebene vergleichbares Verfahren beschreiben Garg et al. [GNRM08]. In dieser Arbeit werden durch das visuelle Interface Mengen positiver und negativer Samples definiert. Dadurch wird eine ternäre Klassifikation $Y = \{positiv, dontcare, negativ\}$ erzeugt, die durch automatische logische Induktion (ILP) in entsprechende Regeln umgesetzt wird. Beiden Verfahren gemeinsam ist, dass sie einen höherdimensionalen Merkmalsraum auf die zwei Bilddimensionen abbilden. Auch wenn eine Visualisierungstechnik nicht alle Attribute eines

Tupels in X darstellen kann, so ist es dennoch möglich, dass das Klassifikationsverfahren für die Modellierung auch solche Merkmale berücksichtigt, die nicht dargestellt werden.
Allgemein eignet sich die interaktive Definition eines Musters als Trainingsdatensatz für ein Klassifikationsverfahren eine sehr flexible Strategie für die intensionale Beschreibung eines Musters. Jede Visualisierungstechnik, die direkte Selektion unterstützt, ist mit dieser Strategie nutzbar. Jedes Analyseverfahren, das die Samples entweder als Trainingsdaten oder auch als Modellparameter (wie etwa bei der *Nearest-Neighbor-Klassifikation*) verwendet ist - abhängig von den Datentypen der Attribute des Merkmalsraums - ebenfalls mit dieser Strategie nutzbar.
Beispielsweise setzen einige Heuristiken für Entscheidungsbäume (z.B. ID3 [Qui87]) eine Diskretisierung der Attribute voraus. Ebenso wie ein Trainingsdatensatz normalerweise nicht die zu klassifizierende Menge vollständig abdeckt, ist es auch nicht notwendig, alle Objekte des Musters interaktiv zu selektieren. Zu beachten ist jedoch, dass Klassifikationsverfahren mit dieser Strategie lediglich eine Beschreibung des Musters liefern können, jedoch nicht ohne weiteres sichergestellt ist, ob diese Beschreibung hinreichend genau ist. Es ist nicht zwangsläufig klar, wie mit dem automatischen Verfahren die Grenzen zwischen den Klassen aus den Samples rekonstruiert und beschrieben werden. Diese Form des Feedbacks wird im zweiten Abschnitt dieses Kapitels allgemein beschrieben.

3.1.4.2 Clustering

Beim Clustering handelt es sich um die Bestimmung einer Funktion $f : X \rightarrow Y$, wobei die Eingabedaten im Gegensatz zur Klassifikation nur Elemente der Grundmange X darstellen. Die Menge Y der zugeordneten Cluster ist a-priori nicht bekannt. Die Identifikation von spezifischen und sinnvollen Clustern abhängig von der Verteilung der Datensätze innerhalb des Merkmalsraums der Daten ist daher das zentrale Problem beim Clustering. Ein Cluster fasst im Idealfall ähnliche und vom Rest abzugrenzende Menge von Datensätzen zusammen. Kritisch beim Clustering ist besonders die Anzahl der Freiheitsgrade, die bestimmen, was „Ähnlichkeit“ formal bedeutet. Im Allgemeinen dürfen die Distanz- bzw. Ähnlichkeitsmaße beim Clustering unabhängig vom Clusteringverfahren gewählt werden. Es handelt sich dabei um komplexe Verfahrensparameter, die vor dem Clustering festgelegt werden. Diese werden im Abschnitt 3.1.4.6 gesondert betrachtet.
Zu den Modellparametern gehört beispielsweise die Anzahl der Cluster. Die Anzahl der Cluster steuert im *K-Means*-Verfahren auch das Verfahren selbst. Beim konventionellen K-Means Algoritmus ist K frei definierbar - die Anzahl der Cluster wird also vorbestimmt. Das Verfahren kann bezüglich des Parameters K als auch bezüglich der Initialisierung der Cluster instabil sein. Wenn das *K-Means* Verfahren Ergebnisse liefern soll, die bezüglich der Initialisierung stabil sein sollen, erfordert dies einen höheren Aufwand [Mir05]. Da die Initialisierung die Selektion von Punkten im n-dimensionalen Raum erfordert, ist es naheliegend, diese manuell durchzuführen. Basierend auf einer oder mehrerer Ansichten der Daten, kann der Nutzer potentielle Cluster identifizieren. Auf vergleichbaren Ansätzen basieren beispielsweise die Arbeiten von Bishop et al. [BT98] und Chen et al. [CL04]. In beiden Ansätzen arbeitet der Nutzer mit einer wählbaren linearen Projektion der Datenpunkte. Durch die Interaktion können Cluster identifiziert, verändert und sukzessive verfeinert werden. Ein

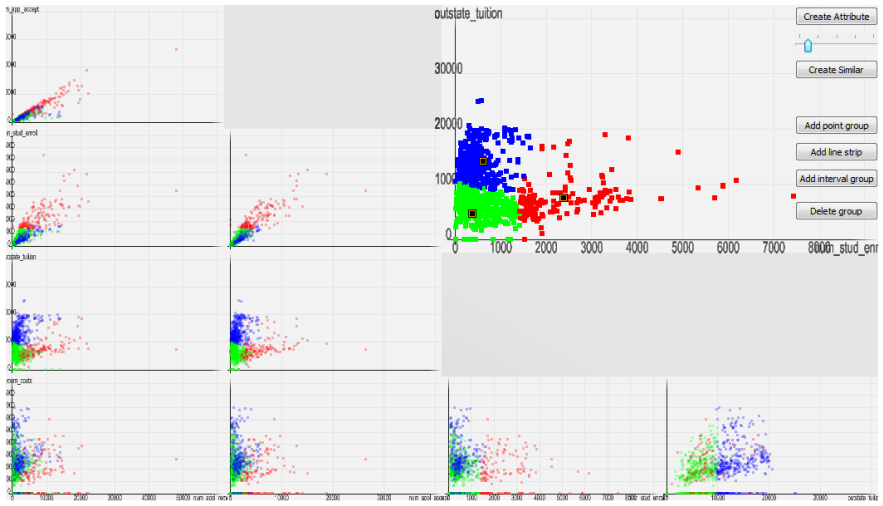


Abbildung 3.8: *In dieser Implementierung einer Scatterplotmatrix kann der Anwender interaktiv Cluster definieren. Der Anwender setzt die Zentroide, wonach die Cluster automatisch über ein „Nearest-Neighbor"-Verfahren berechnet werden. Die Färbung in den Diagrammen folgt dann nach diesem Schema, auch wenn die Zentroide bewegt werden. Beim manuellen Clustering kann die Grenze zwischen Clustering und Klassifikation nicht mehr scharf gezogen werden.*

Problem beim Clustering besteht darin, dass die Cluster und deren Anzahl durch die Verteilung der Datenpunkte auf sehr unterschiedlichen Skalengrößen charakterisiert sein könnten. Durch die gekoppelte Untersuchung der Daten mit Visualisierungstechniken und Clusteringverfahren lässt sich das Problem für die Identifikation der Cluster weitgehend entschärfen.
Bei der interaktiven Definition von Clustern durch den Anwender (siehe Abbildung 3.8) stellt sich jedoch die Frage, inwiefern man noch von Clustering in Abgrenzung zu Klassifikationsverfahren sprechen darf. Durch die Kopplung wird das Clusteringproblem in zwei Teilaufgaben zerlegt. Die erste Teilaufgabe ist die Identifikation von Clustern, die Zweite ist die Beschreibung bzw. Modellierung der Cluster. Clustering und Klassifikation lassen sich anhand der ersten Teilaufgabe voneinander abgrenzen. Bei der Separation von Mustererkennung und Musterbeschreibung findet diese Abgrenzung also bei der Aufgabe statt, die dem menschlichen Anwender obliegt.
Die Aufgaben Klassifikation und Clustering lassen sich daher in der Interaktion voneinander abgrenzen. Die Aufgabe bei der Klassifikation besteht darin, jene Datenmuster, die aus der vorgegebenen Klassifikation folgen, zu identifizieren. Insbesondere kann der Anwender dabei eine Separation von Muster und Rauschen vornehmen. Für die Klassifikation relevant ist daher das visuelle Attribut, auf das die Menge Y abgebildet wird. Im Unterschied dazu gibt es beim Clustering keine Vorgaben über die Relevanz bestimmter visueller Attribute. Da Y bei der interaktiven Klassifikation aus technischer Sicht wie ein beliebiges anderes Datenattribut des entsprechenden Skalentyps behandelt werden kann, kann eine Visualisierung gleichermaßen für beide Aufgaben verwendet werden.

3.1.4.3 Auswahl von Verfahrensparametern

Ansätze, in denen Verfahrensparameter visualisiert werden, sind vergleichsweise selten. Ein Beispiel dafür die die sogenannten *Receiver Operating Characteristics* [BN01], in denen Gütemaße für die Klassifikation einander gegenüber gestellt werden können. Bei weitem häufiger werden die Verfahrensparameter über das GUI gesteuert. Die Steuerung der Verfahrensparameter über das GUI hat jedoch den Nachteil, dass jeder Verfahrensparameter isoliert betrachtet und gesteuert wird: Die Steuerungsmetaphern schaffen keinen Kontext, der die Kriterien und deren Bezüge für die Wahl eines Parameterwertes abbildet.
Dies ist dann akzeptabel, wenn die Kriterien für oder wider die Wahl eines Verfahrensparameterwertes bekannt und hinreichend einfach sind. Definiert die Problemstellung der Analyse eine Vorgabe, dass ein Klassifikator eine bestimmte Fehlerrate für einen Fehler 1. Art nicht überschreiten darf, dann kann ein Parameter, der die Komplexität des Verfahrens steuert, auch über das GUI entsprechend angepasst werden.
Von dieser Voraussetzung kann man jedoch nicht in jedem Fall ausgehen: Sobald ein Verfahrensparameter von mehreren Informationen in komplexer Weise abhängt, und erst recht dann, wenn die Kriterien für die Wahl eines Parameters überhaupt nicht bekannt sind, muss ein Kontext konstruiert werden, der diese Kriterien geeignet zusammenfasst. Wie in Abschnitt 2.1 dargelegt, darf die Wahl eines Verfahrensparameters nicht willkürlich erfolgen.
Potentiell relevante Kriterien für die Wahl eines Verfahrensparameters sind beispielsweise:

- das Schema der zu verarbeitenden Daten,
- Vorwissen über die Daten,
- Historie der Analyse,
- bekannte Wechselwirkungen zwischen mehreren Verfahrensparametern.

Das Datenschema ist im Unterschied zu den Datenstrukturen zu Beginn der Analyse bekannt und beschreibt die trivialen Beziehungen zwischen Entitäten und Attributen. Vorwissen über die Daten über die schließt dabei sowohl innere Bezüge der analysierten Daten ein, als auch bekannte Beziehungen zwischen den Daten und der Problemstellung der Analyse.
Die Historie der Analyse umfasst sowohl während der Analyse neu gefundene Informationen, als auch das Wissen darüber, welche Parameterwerte in vergleichbaren Situationen bereits „gute“ Ergebnisse geliefert haben. Die Exponierung von Alternativen und die Begründbarkeit der Wahl von Verfahrensparametern ist die Vorraussetzung für eine methodische Analyse.
Bezüglich der Art der Informationen, die dargestellt wird, und aus denen der Anwender wählen kann, unterscheidet sich diese Kopplung von anderen Kopplungsvarianten. Gegenstand der Untersuchung ist nicht der Merkmalsraum der Daten, sondern in gewisser Weise der Merkmalsraum, den die Freiheitsgrade für die Steuerung eines Verfahrens bilden. Untersucht wird dabei die Menge aller möglichen Konfigurationen von Verfahrensparametern und deren Bezüge zu anderen innerhalb der Analyse nutzbaren Informationen.
Interessant ist diese Kopplung deshalb, weil sie sich bezüglich der *Techniken*, die dabei eingesetzt werden, nicht prinzipiell von den Kopplungen unterscheidet, die direkt auf den Daten

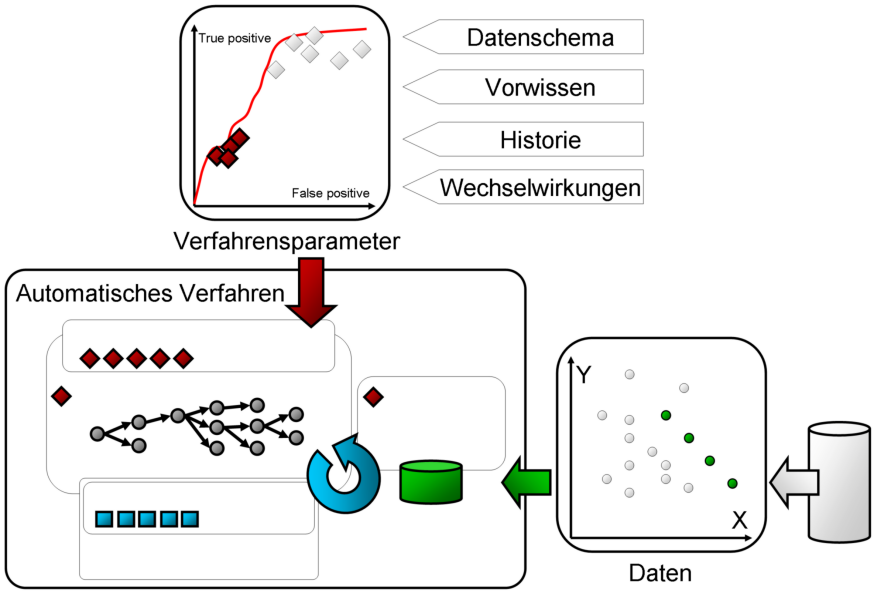


Abbildung 3.9: *Die Wahl eines Verfahrensparameters muss begründet werden. Eine Kopplung zwischen visuell-interaktiven und automatischen Verfahren, die an den Verfahrensparametern ansetzt, kann die Optionen und Kriterien bei der Wahl eines Verfahrensparameters exponieren. Relevant für diese Entscheidung können beispielsweise Datenschemata, Vorwissen über die Daten oder die Historie der Analyse sein. Eine Visualisierung ist die Möglichkeit, diese Kriterien miteinander auf die gleiche Weise in Bezug zu setzen, wie die Daten selbst (unten rechts). Die Visualisierung böte beispielsweise die Möglichkeit, alle bisher gewählten Parameterwerte mit den Gütekriterien eines Verfahrens in Beziehung zu setzen.*

operieren. Die Vorraussetzung dafür, dass überhaupt methodisch Parameter gewählt werden können, ist, dass die Beziehung zwischen der Wahl eines beliebigen Parameters und einer (nach gegebenen Gesichtspunkten) erfolgreichen Analyse gewissen Regeln unterliegt. Die Methoden, mit denen in der explorativen Datenanalyse Regelmäßigkeiten in den Daten gefunden werden sollen, lassen sich mithin auch nutzen, um Regelmäßigkeiten bei der Wahl der Parameter zu finden und damit die Regeln, nach denen „gute“ Parameterwerte gewählt werden können.

Diese Variante der Kopplung überschneidet sich mit dem Grundgedanken der Separation von Mustererkennung und Musterbeschreibung. Die Komplexität der Entscheidung wird durch die Visualisierung potentieller Kriterien zumindestens teilweise auf die Fähigkeiten zur Mustererkennung übertragen. Die Interaktion in dieser Visualisierung dient dabei in erster Linie der Wahl der Parameter. Eine Musterbeschreibung wird dann relevant für die Beschreibung der Methodik der Analyse.

Die Wahl von Attributen eines Datensatzes für ein Verfahren ist ein Spezialfall bei der Wahl von Verfahrensparametern. Eine Möglichkeit, durch eine Visualisierung einen Kontext für diese Wahl herzustellen wird im folgenden Abschnitt beschrieben (siehe Abbildung 3.9).

3.1.4.4 Auswahl von Attributen

Die Auswahl von Attributen ist einer der weitreichendsten Teilprozesse innerhalb der Analyse. Langley [Lan94] bezeichnet die Separation relevanter und irrelevanter Attribute als eines der zentralen Probleme des Machine-Learning. Es handelt es sich im Prinzip um einen Spezialfall der Dimensionsreduktion. Im Machine-Learning wird *Feature Selection* von *Feature Extraktion* unterschieden. Bei der Feature Extraktion können synthetische Attribute erzeugt werden, die darauf optimiert werden, dass sie Information aus möglichst vielen ursprünglichen Attributen zusammenfassen.
Ausgangssituation ist dabei ein Merkmalsraum, der mehr Attribute enthält, als in den in der Analyse nachfolgenden Teilprozessen weiterverarbeitet werden können. Gesucht wird eine Teilmenge der Attribute, die nicht größer ist als vorgegeben, und die dennoch möglichst viele Informationen über den ganzen Merkmalsraum zusammenfassen. Wenn nicht garantiert ist werden kann, dass diese Informationen auch nach der Auswahl noch für die Weiterverarbeitung zur Verfügung stehen, geht man das Risiko ein, dass die Ergebnisse in erster Linie ein Artefakt dieser Auswahl sind.
Eine Vorauswahl von Attributen muss für viele automatische Verfahren, für *alle* Visualisierungstechniken und natürlich auch dann getroffen werden, wenn der der Anwender die Ergebnisse direkt verwendet. Beispielsweise gibt es automatische Verfahren, die mit der Dimension des Merkmalsraums schlecht skalieren. Dies trifft beispielsweise auf einfaches *K-Means-Clustering* [Mir05] zu. Die Ursache besteht grob formuliert darin, dass die mit zunehmender Dimensionalität des Merkmalsraums die Verteilungsdichte der Daten im Merkmalsraum eventuell so klein wird, bis sich die Struktur des Datensatzes sich kaum noch von einer zufälligen Verteilung abhebt [JDM00]. Schlecht skalierende Verfahren können dann instabil werden bezüglich der Wahl der Trainingsdaten.
John et al. [JKP94] beschreiben ein Beispiel für das Training eines Entscheidungsbaums, in dem Abhängigkeiten in den Attributen Artefakte erzeugen. Die Auswahl der Attribute muss auch in diesem Fall die bekannten Abhängigkeiten berücksichtigen.
Das Gros der Literatur, die sich mit „Visualisierung“ in Verbindung mit „Attributauswahl“ auseinandersetzt, bezieht sich auf die „Attributauswahl für die Visualisierung“ anstatt auf „Visualisierung für die Attributauswahl“. Der Grund dafür liegt auf der Hand: Die Skalierbarkeit bezüglich der Dimension des Merkmalsraums ist eine der fundamentalen Herausforderungen der Visualisierung insbesondere für die explorative Datenanalyse.
Für jede Kombination von Ausgabegerät und Visualisierungstechnik gibt es eine nicht überschreitbare obere Grenze der darstellbaren Dimensionen, die durch den Größe des Fokus der Aufmerksamkeit des Menschen noch weiter eingeschränkt wird [War04b]. Daraus folgt, dass jedes Ergebnis und jedes Muster, das mit einer Visualisierungstechnik gefunden wird, mittelbar abhängig ist von dieser Auswahl.
Die Idee, die Auswahl von Attributen im allgemeinen Fall von Visualisierungstechniken abhängig zu machen, scheint daher das Grundproblem nur zu verlagern: Die Attribute, die auf der Basis einer Visualisierung gewählt werden könnten, wird eine Teilmenge derjenigen sein, die vorher für die Visualisierung ausgewählt wurde.
Anstatt die Auswahl von Attributen über die Kopplung zwischen Mustererkennung und Musterbeschreibung zu betreiben, bieten sich zwei Alternativen an:

- Manuelle Auswahl von Attributen durch den Anwender über ihre Namen. Dies ist nur möglich wenn der Bezug der Attribute zur Anwendungsdomäne durch ihre Namen erkennbar ist.

- Verwendung automatischer Verfahren, die bezüglich der Dimension des Merkmalsraums gut skalieren.

Jedoch sind die beiden anderen Strategien ebenfalls nicht unproblematisch. Im ersten Fall hängt die Auswahl vom Vorwissen des Anwenders ab. Es handelt sich zwar die wahrscheinlich am häufigsten verwendete Strategie, jedoch steht sie dem Anspruch von Visual Analytics - *Discover the Unexpected* diametral gegenüber. Viel besser eignet sich die manuelle Auswahl von Attributen für die konfirmative Datenanalyse.
Bei der automatischen Auswahl wird man in Reinform mit dem grundsätzlichen Problem konfrontiert, dass die Ergebnisse ebenso ein Artefakt der Daten, wie auch ein Artefakt der verwendeten Verfahren und Parameter sein können. Eine Strategie bei der automatischen Auswahl der Attribute ist das systematische Testen verschiedener Selektionen der gewünschten Größe, auf die wiederum Clustering oder Klassifikationsverfahren angewendet werden. Da die Anzahl möglicher Kombinationen sehr hoch sein kann, müssen geeignete Heuristiken dafür verwendet werden. In [Lan94] werden dafür beispielsweise Genetische Algorithmen eingesetzt. Um die Tests zu bewerten und zu optimieren, werden entsprechende Gütefunktionen eingesetzt. Dadurch überträgt sich der Bias, der durch die Auswahl von Verfahren und Parameter erzeugt wird, überträgt sich auf die automatische Auswahl.
Daraus ergibt sich, dass auch bei automatischen Verfahren das Problem für die Attributs-

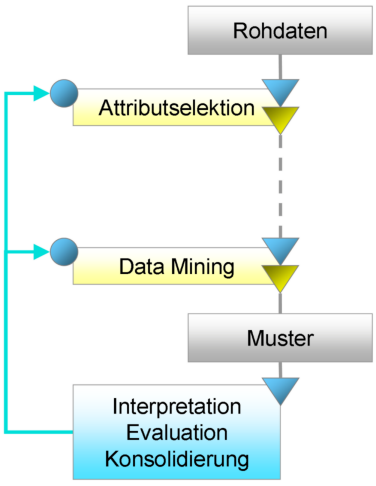


Abbildung 3.10: *Eine explorative Analyse ist stets ein iterativer Prozess. Die Auswahl von Attributen und die nachfolgenden Data-Mining Schritte sind wechselwirkende Prozesse. Im Allgemeinen kann man nicht davon ausgehen, dass die erste Auswahl alle potentiell relevanten Attribute berücksichtigt. Durch die Analyse vergrößert sich das Wissen um die Beziehungen in den Daten. Dieses Wissen muss wiederum im Auswahlprozess für die folgenden Iterationen genutzt werden. Häufig wird die Auswahl von Attributen direkt durch den Anwender getroffen. Wenn stattdessen automatische Verfahren für die Selektion verwendet werden, verschiebt sich das Problem der Attributauswahl auf das Problem der Auswahl eines geeigneten automatischen Verfahrens.*

elektion ebenfalls verlagert wird - in diesem Fall auf die Selektion des richtigen Verfahrens, der Parameter, der Distanz- und Gütefunktionen. Dies ist unter dem Gesichtspunkt problematisch, dass die Auswahl über lesbare, symbolische Informationen (d.h. die Namen der Attribute) ersetzt wird durch die aus Auswahl von subsymbolische Parametern für die automatischen Verfahren.
Deutlich wird eine Wechselwirkung zwischen jenen Analyseschritten, die die eigentlichen Verfahren zur Mustererkennung nur vorbereiten sollen, und dem Schritt der Mustererkennung selbst. Mit den nachfolgenden Schritten werden genau die Informationen gefunden, die man für die Auswahl der Attribute am besten nutzen könnte. Es handelt sich dennoch nicht um ein „Henne-Ei“ Problem:
Methoden für die Auswahl von Attributen basieren auf der Bestimmung von Ähnlichkeiten oder Abhängigkeiten. Auf diese Weise können beispielsweise Attribute gefunden werden, die mit einem minimalen Verlust von Informationen für die folgenden Schritte der Analyse unberücksichtigt bleiben können. Im iterativen Prozess wird das Wissen um Ähnlichkeiten und Abhängigkeiten schrittweise verfeinert. Man kann davon ausgehen, dass jede bereits gefundene Beziehung zwischen Attributen eines Merkmalsraums die Grundlage für die Attributauswahl verfeinert. Dies trifft natürlich auch auf Beziehungen zu, die nur wenige Attribute betreffen.
Das bedeutet, dass jedes Verfahren - unabhängig davon, wie gut es mit der Anzahl der Attribute skaliert - indirekt genutzt werden kann, um die Auswahl der Attribute, in den folgenden Schritten zu verbessern. Dafür müssen jedoch auch alle Informationen, die für die Auswahl der Attribute potentiell relevant sind, in einem gemeinsamen Zusammenhang gestellt werden.
Hier wird eine für die Auswahl von Attributen eine Strategie vorgeschlagen, in der visuell-

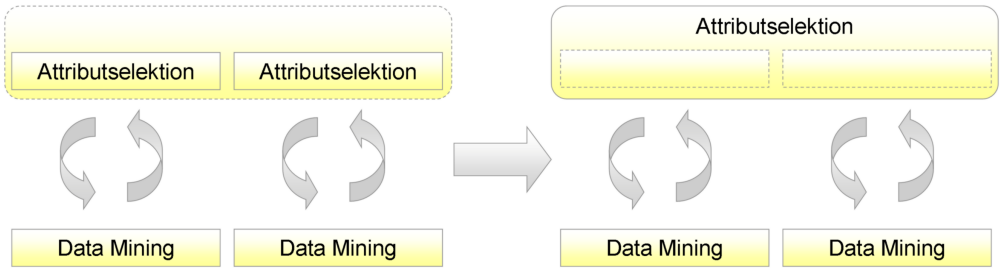


Abbildung 3.11: *Jedes Analyseverfahren kann die Informationsgrundlage für die Auswahl der Attribute im nächsten Schritt der Analyse verbessern. Dies gilt auch dann, wenn ein Verfahren nicht alle Attribute eines Datensatzes berücksichtigt. Da jedes einzelne Verfahren jedoch stets nur einen Teil der Informationen beitragen kann, müssen die Informationen bewertet und konsolidiert werden. In den meisten Fällen ist dies Aufgabe des Nutzers. In der hier vorgeschlagenen Strategie unterstützt eine Visualisierung der Informationen diese Konsolidierung, bei der alle bisherigen Teilergebnisse für die nächste Auswahl genutzt werden. Diese Strategie wird in Abschnitt 4.2 vorgestellt.*

interaktive und automatische Verfahren gekoppelt werden. Der Ansatzpunkt für die Kopplung ist dabei die Auswahl der Attribute für die Analyseverfahren. Die Visualisierung unterstützt dabei die Konsolidierung eines Teils der Informationen, die für die Attributselektion notwendig sind.
Bei dieser Konsolidierung ist die Information über alle bereits bekannten oder gefundenen

Beziehungen wichtiger als ein einzelnen automatisches Verfahren. Verantwortlich für die systematische Untersuchung des Merkmalsraums und aller potentiellen Beziehungen ist dabei der Anwender. Die Auswahl von Attributen erfolgt daher interaktiv, und eine Grundlage für die Entscheidung ist die Visualisierung der bereits gefundenen Beziehungen zwischen Attributen. Die Strategie ist ein zyklischer iterativer Prozess, der folgende Schritte umfasst:

- Auswahl der Attribute durch den Anwender
- Durchführung eines automatischen oder visuell-interaktiven Verfahrens für die Suche neuen Beziehungen
- Entscheidung des Anwenders über die Relevanz der neu gefundenen Beziehungen
- Visualisierung der gefundenen Attributbeziehungen

Die Analyse kann dabei sowohl mit der Auswahl von Attributen, mit der Durchführung eines automatischen Verfahrens auf allen Attributen oder auch mit der Visualisierung bereits bekannter Beziehungen beginnen. Kernpunkt der Strategie ist die Gegenüberstellung zwei verschiedener Sichtweisen auf die Beziehungen zwischen Attributen.
Eine Sichtweise ist die Darstellung der Beziehung als Muster oder Strukturen, die sich in der Darstellung der Daten manifestieren. Die zweite Sichtweise ist die Darstellung von Attributen und Beziehungen als Entitäten eigenen Rechts. Die Strategie folgt in dieser Hinsicht dem Ansatz von Yang et al. [YXRW07]. Yang schlägt ein System für das Management von *Nuggets* vor, die genau die Muster bezeichnen, die ein oder mehrere Anwender als potentiell relevant identifiziert haben. Ziel beim Management der Nuggets ist die Konsolidierung der durch sie repräsentierten Informationen. Diese Konsolidierung wird durch eine Analyse der Ähnlichkeiten der durch die Muster repräsentierten Mengen durchgeführt.
Im Gegensatz zum Ansatz von Ward, der auf einen inhaltlichen Vergleich gefundener Muster abzielt, abstrahiert die hier vorgestellte Strategie noch stärker von Merkmalsraum und fokussiert in erster Linie nur auf die Tatsache, *dass* innerhalb einer bestimmten Teilmenge von Attributen ein Muster bzw. eine Beziehung gefunden wurde. Insbesondere ist es dabei unerheblich, mit welchen Verfahren eine Beziehung gefunden wurde. Die Strategie für die Suche nach Attributen stützt sich auf die Annahme, dass jedes Verfahren spezifische Stärken bei der Suche nach Mustern hat, und infolgedessen mit jedem Verfahren spezifische Informationen für die bessere Beschreibung des Merkmalsraums gefunden werden können. Darüber hinaus bleibt Information über den Merkmalsraum, nachdem sie einmal gefunden wurde, auch im Fokus des Anwenders. Daher können die Anforderungen an die Verfahren für die Analyse und Attributauswahl abgeschwächt werden:

- Die Verfahren müssen nicht den vollständigen Merkmalsraum untersuchen.
- Bei der Beschreibung einer Beziehung müssen nicht alle relevanten Attribute berücksichtigt werden.

Insbesondere folgt aus dieser Abschwächung, dass auch solche Verfahren zu dieser Verfeinerung beitragen können, die spezialisiert sind auf bestimmte Datentypen, Modellfamilien

und vor allem beschränkt sind in der Anzahl der Attribute, die gleichzeitig in Bezug gesetzt werden können. Dies erweitert das Repertoire an Verfahren, die in diesem Prozess genutzt werden können erheblich, und schließt insbesondere auch Visualisierungstechniken mit ein.
Ebenfalls folgt aus dieser Abschwächung, dass keineswegs nur prägnante Muster oder Beziehungen in die Darstellung eingeschlossen werden sollen. Die jeder Entscheidung des Analysten zugrundeliegende Annahme muss ja darin bestehen, dass sie potentiell falsch oder mindestens nicht die Beste ist. Die Darstellung der gefundenen Beziehungen dient daher nicht nur der Suche nach neuen, sondern auch nach der Verbesserung, Verfeinerung oder Verallgemeinerung bekannter Beziehungen mit anderen Verfahren oder Konfigurationen.
In der hier vorgestellten Abstraktion bilden die Attribute ein Netzwerk bzw. einen Hypergraphen, das mit entsprechenden Methoden darstellt werden kann. Dies entspricht dem gleichen Abstraktionsprinzip mit dem beispielsweise auch Bayessche Netze dargestellt werden. Im Unterschied zu Bayesschen Netzen repräsentieren die Kanten jedoch keine Informationen, die spezifisch wären für ein bestimmtes analytisches Modell. In Verbindung mit der Strategie der iterativen Verfeinerung kann das Problem der Attributauswahl auf ein Problem der Netzwerkanalyse transformiert werden.
Dies schreibt zunächst nicht vor, ob dieses Problem mit automatischen oder visuell-interaktiven Verfahren gelöst wird. Innerhalb dieses Konzepts wurde damit jedoch eine Möglichkeit erschlossen, mit der die Auswahl von Attributen durch Visualisierungstechniken (in diesem Fall für Graphen) unterstützt werden kann. Die Auswahl von Attributen für ein Verfahren entspricht demnach der Selektion von Knoten oder auch Kanten innerhalb dieses Graphen. Jedes beschriebene neue Modell erzeugt eine neue Menge von Kanten. Die Strategie für die explorative Analyse wird übertragen auf das Zusammenfügen von Elementen eines Graphen (bzw. eines Waldes).
Die Auswahl der Attribute durch den Nutzer stellt eigentlich eine kognitive Leistung dar. Durch die Visualisierung der dafür relevanten Informationen wird das mentale Bild exponiert, mit dem sich der Anwender einen Überblick über die gesamte Analyse machen muss. Die Aufgabe der Attributauswahl wird damit teilweise zu einer Aufgabe der Strukturerkennung in Graphen.

Dass diese Strategie umgesetzt werden kann, wird im folgenden Kapitel nachgewiesen (siehe Abschnitt 4.2). Darin wird eine Visualisierungstechnik für die Attribute und die bereits gefundenen Beziehungen vorgestellt. Mit dieser Visualisierung werden die Attribute für die Verfahren ausgewählt, die für die Modellierung neuer oder verfeinerter Beziehungen vorgesehen sind. Durch die Übersicht über die bereits identifizierten Ähnlichkeiten und Abhängigkeiten kann der Anwender gezielt solche Attributmengen identifizieren, mit denen eines der vier oben genannten Ziele umgesetzt werden kann.
Die Forschung darüber, wie mit Visualisierungstechniken die Attributauswahl gesteuert und verfeinert werden kann, wird in vergleichsweise wenigen Arbeiten behandelt. Ein Beispiel dafür sind die Arbeiten von Jankun-Kelly et al. [JK03, JKMG07]. Diese Arbeiten behandeln interne und externe Repräsentierungen des Parameterraums von Visualisierungstechniken. Externe Repräsentierungen sind dabei selbst Visualisierungen, die für die Auswahl der Parameter genutzt werden. Der Parameterraum für die Steuerung der Visualisierung wird dabei vollständig abgedeckt, beschreibt also inbesondere auch die Attribute, die dargestellt werden. Jankun-Kelly beschreibt dabei in [JK03] ein *Spreadsheet-Interface* für die Steuerung der Visualisierung. Wichtigstes Ziel ist jedoch die effektive Suche nach „guten“ Parame-

terkonfigurationen der Visualisierung. Die Auswahl der dargestellten Attribute aus einem Merkmalsraum, die im Prinzip ebenfalls zu diesen Parametern gehören, liegt jedoch nicht im Fokus dieser Arbeiten.

3.1.4.5 Dimensionsreduktion und Merkmalsextraktion

Dimensionsreduktion bzw. Merkmalsextraktion kann als Verallgemeinerung der Attributauswahl verstanden werden. Bei der Attributauswahl werden die Merkmale der Datensätze unverändert übernommen; es bestehen lediglich die Optionen ein Attribut für die Analyse zu verwenden oder nicht. Bei der Dimensionsreduktion ist es möglich, aus der Kombination von Merkmalen synthetische Attribute zu bilden. Synthetische Attribute fassen die Information mehrere Attribute über die Verteilung der Datenobjekte im Merkmalsraum zusammen und stellen hinsichtlich der Informationsdichte eine effektivere Repräsentierung des Merkmalsraums dar.
Auch hier muss deutlich unterschieden werden zwischen Arbeiten zum Thema „Dimensionsreduktion für die Visualisierung“ und zum Thema „Visualisierung für Dimensionsreduktion“. Das Thema „Dimensionsreduktion für die Visualisierung“ wird in der Literatur ausführlich behandelt. Dies liegt nicht zuletzt daran, dass einige Visualisierungstechniken nicht nur eine Attributselektion, sondern allgemein eine Dimensionsreduktion voraussetzen (siehe Abschnitt 2.3.5.4). Die Visualisierung ist dann das Ergebnis einer Projektion, deren Parameter automatisch bestimmt wurden.
Lässt man stattdessen zu, dass die Parameter der Projektion interaktiv über die Visualisierung bestimmt werden, bietet das im Umkehrschluss eine Strategie für die interaktive Bestimmung synthetischer Attribute. Vergleichsweise häufig ist die Darstellung von Punktwolken im dreidimensionalen Raum als lineare Projektion. Durch interaktive Kontrolle der drei Rotationsachsen kann der Nutzer die Projektion bestimmen die - dem Augenschein nach - die Strukturen der Daten am besten wiedergibt. Die beiden Achsen der Visualisierung lassen sich dann als synthetische Attribute interpretieren.
Was mit drei Dimensionen noch funktioniert, scheitert bei mehr Dimensionen daran, dass der Anwender den Effekt einer Rotation nicht mehr nutzen kann, um die Punkteverteilung im Raum zu rekonstruieren; mithin hat er keine natürliche Kontrolle über einzelne Rotationsachsen. Faith [Fai07] beschreibt einen anderen Ansatz für die implizite Definition einer Rotation im n-dimensionalen Raum. Der Nutzer bestimmt eine beliebige Punktmenge in der Darstellung durch direkte Selektion. Auf der Basis dieser Punktmenge wird die Projektion bestimmt, die diese Punktmenge maximal separiert und die Visualisierung wird in diese Projektion überführt. Hintergrund dieses Ansatzes ist das Problem, dass aus nur einem Blickwinkel nie beurteilt werden kann, ob ein Cluster von Datensätzen auch im hochdimensionalen Raum eine zusammenhängende Struktur bilden. Ist die Struktur dagegen auch nach der Projektion noch als Cluster erkennbar, kann man davon ausgehen, dass sie auch durch keine andere Projektion separiert wird und die selektierten Datensätze tatsächlich einen Cluster bilden.
Wie ähnliche Konzepte für die Dimensionsreduktion auch für Visualisierungstechniken angewendet werden können, die mehr als zwei Dimensionen darstellen können, zeigen Ward

et al. [WLT94]. Sie beschreiben mehrere Strategien für die Dimensionsreduktion von Attributen für eine Visualisierung basierend auf der *Dimensional Stacking*-Technik, darunter auch eine Rotation des Merkmalsraums für die Bestimmung synthetischer Dimensionen. Im Unterschied zur Darstellung von Punktwolken findet diese Rotation jedoch *vor* der visuellen Abbildung statt. Im Bild nimmt man daher nicht die Rotation wahr, sondern die Umverteilung der Daten entlang der synthetischen Dimensionen.

3.1.4.6 Distanzfunktionen

Die Bestimmung von Distanzfunktionen für ihren Einsatz in verschiedenen Suchverfahren steht seltener im Fokus als die eigentlichen Heuristiken. Distanzmaße auf dem Merkmalsraum haben eine zentrale Bedeutung insbesondere beim Clustering und der Untersuchung von Ausreißern. Die Qualitätsfunktion, anhand derer eine Heuristik optimiert wird stellt einen zusätzlichen Freiheitsgrad bei der Konstruktion der Verfahren dar.
Die Distanzfunktion auf einem Merkmalsraum M ist eine Funktion $dist : M \times M \rightarrow \Re$. Man kann leicht zeigen, dass die Bestimmung einer geeigneten Distanzfunktion mindestens so schwierig ist wie das Problem der Dimensionsreduktion. Eine Dimensionsreduktion kann als Konstruktion einer Distanzfunktion formuliert werden, bei der zusammenfallende Punkte des Merkmalsraums Äquivalenzklassen bilden, innerhalb deren die Distanz 0 ist. Gleichzeitig bestimmt die Distanzfunktion aber in hohem Maße das Verhalten der Suchverfahren und damit das Ergebnis der Suche. Ist die Distanzfunktion frei wählbar, ist es möglich zu jeder Punktmenge $X \subset M$ und jedem gegebenen oder gewünschten Clustering $clust : X \rightarrow C$ auf X eine Distanzfunktion auf dem Merkmalsraum anzugeben, die auf X dieses Clustering erzeugt. In diesem Extremfall wäre das Ergebnis der Analyse nicht das Clustering, sondern das danach konstruierte Distanzmaß. In Abschnitt 2.3.4 wurde bereits dargelegt, dass in jeder Analyse, in der Distanzfunktionen frei wählbar oder gar parametrisierbar sind, am Ergebnis bestimmt werden muss, welcher Teil des ein Artefakt der Daten und welcher Teil ein Artefakt der Distanzfunktion ist.
Eine Distanzfunktion gilt im Allgemeinen nicht als das Ergebnis, sondern die Grundlage für eine automatische Analyse. In vielen Fällen existiert ein Repertoire verschiedener Distanzfunktionen (siehe 2.3.4), die auf den Verfahren und Daten getestet werden können. Verschiedene Möglichkeiten für die Definition von Distanzfunktionen lassen sich aus der Tatsache ableiten, dass Dimensionsreduktion und Attributauswahl als Sonderfälle der Konstruktion von Distanzfunktionen betrachtet werden können: Eine Visualisierung, die auf einer Dimensionsreduktion des Merkmalsraums basiert, kann eine Distanzfunktion beschreiben. Voraussetzung ist jedoch, dass ausschließlich der Nutzer die Parameter der Visualisierung verändert, bis er Strukturen in den Daten erkennt. Würde die Dimensionsreduktion hingegen automatisch durchgeführt, dann wären die Parameter der Visualisierung umgekehrt ein Artefakt der Distanzfunktion des verwendeten Verfahrens.
Schreck et al. [SBvLK09] beschreiben ein Clusteringverfahren für mehrdimensionale Datenvektoren basierend auf einer zweidimensionalen Kohonenkarte (*SOM*, siehe Abschnitt 2.3.5.4). Die Karte wird auf den Bildraum projiziert. Da die Kohonenkarte ein nichtlineares Projektionsverfahren darstellt, wird die Korrespondenz zwischen Merkmalsraum und

Bildraum durch die Darstellung von Repräsentanten hergestellt. Die lokale Dichte an Datenpunkten in der Projektion dient unter anderem als Indikator für die Qualität der Projektion. Ein besonderes Merkmal dieses Ansatzes besteht darin, dass die Projektion durch die Fixierung bestimmter Repräsentanten im Bildraum gesteuert werden kann. Vor oder auch während der Iterationen kann ein Nutzer Punkte des Bildraums festlegen, auf die bestimmte Punkte des Merkmalsraums abgebildet werden müssen. Bezogen auf die Bestimmung eines Distanzmaßes, bestimmt der Nutzer also in gewissen Grenzen, welche Merkmalsgruppen, in welchem Bereich der Projektion hoch aufgelöst werden und welche nicht. Auch wenn dabei das Abstandsmaß auf den Merkmalsraum nicht direkt definiert wird, kann das Ergebnis der nicht-linearen Projektion verwendet werden. Wie im auch im linearen Fall induziert das Abstandsmaß auf dem projizierten Raum - in diesem Fall der Ebene des Bildes - auch ein Abstandsmaß auf dem Merkmalsraum. Auch hier gilt, dass der Anwender einen hinreichend großen Einfluß auf die Gestaltung der Projektion haben muss.
Das Beispiel von Schreck et al. *(ebd.)* unterscheidet sich nicht nur in der Art der Projektion von den im vorigen Abschnitt genannten Strategien. Anstatt die Projektion zu direkt oder indirekt (wie bei Faith [Fai07]) zu steuern, wird nur das vom Nutzer gewünschte Ergebnis der Projektion durch die Repräsentanten vorgegeben. Der Anwender reproduziert „seine“ Ordnungsstruktur auf dem Merkmalsraum durch die Positionierung im Bildraum, so dass die relative Position die Ähnlichkeit der Repräsentanten beschreibt. Diese Strategie ist in solchen Fällen sinnvoll, in denen der Anwender eine klare Vorstellung von der Bedeutung der Datenobjekte hat, jedoch daraus keine Rückschlüsse auf die Bedeutung einzelner Merkmale oder Attribute machen kann.
Im Gegensatz zu den anderen hier vorgestellten Verfahren ist die Grundlage der Interaktion kein *wahrgenommenes* Muster, sondern eine Ordnungsstruktur im Kopf des Anwenders, die durch die Interaktion mit den Daten erst externalisiert wird: Das Muster wird durch den Nutzer in der Visualisierung konstruiert. Die Komplexität dieser Aufgabe hängt davon ab, wie groß die Korrespondenz zwischen dem mentalen und dem sichtbaren Bild ist und ob die jeweilige Visualisierung für diese Aufgabe überhaupt geeignet ist.

3.1.5 Modellvisualisierung

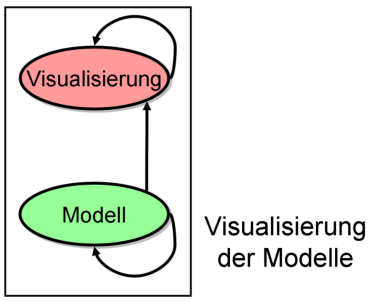


Abbildung 3.12: *Die Modellvisualisierung ist keine neue Kopplung zwischen Visualisierung und Data-Mining Techniken. Sie gehört jedoch deshalb zum Konzept, weil sie die Voraussetzung für die Präsentation der Ergebnisse und für den visuellen Abgleich konstruierter und gegebener Modelle ist.*

Die Modellvisualisierung ist der letzte Schritt des Prozesses der analytischen Transformationen elementarer Artefakte in Modelle. Außer in dem Fall, dass ein prädiktives Modell ausschließlich für die automatische Verarbeitung von Daten eingesetzt werden soll, ist die Präsentation des Modells von ebenso zentraler Bedeutung wie die vorangegangen Schritte. Die Darstellung der Modelle schließt sich an die formale Beschreibung der Muster an.
Die Visualisierung von Modellen soll im Rahmen dieses Konzepts nicht als neue Form der Kopplung zwischen automatischen und visuell-interaktiven Methoden verstanden werden: Für praktisch jedes analytische Modell existiert auch eine visuelle Form der Darstellung. Da die Aufgabe der Modellierung im Rahmen dieses Konzepts jedoch auf die automatischen Verfahren übertragen wird, wird hiermit eine Möglichkeit vorgesehen, mit der das in dem Modell repräsentierte Wissen dem Menschen zugänglich gemacht wird.
Ein Beispiel für die Visualisierung von Assoziationsregeln liefern Blanchard et al. [BPKG07]. Keogh und Lin stellen einen Suffix-Baum als Ergebnis einer Zeitreihenanalyse [LKL05] als Modell dar. Bäume werden ebenfalls bei der Darstellung der Ergebnisse von hierarchischen Clusteringverfahren genutzt (siehe u.a. bei Wilkinson und Friendly [WF09], Eun et al. [NHM+07] sowie Rasmussen und Karypis [RK04]).
Die Darstellung von Entscheidungsbäumen ergibt sich direkt aus der Metapher des analytischen Modells. Auch hierfür existieren zahlreiche Beispiele, wie etwa Ankerst et al. [AEK00], Ho et al [HNN02] oder Ware [WFH+01]. Eine kanonische visuelle Metapher haben beispielsweise auch probabilistische Modelle wie Markovketten oder Bayessche Netze.
Selbst subsymbolische analytische Modelle wurden bereits visualisiert. Ein Beispiel dafür ist der Ansatz von Sniezynski et al. [LPWH06], die Neuronale Netze darstellen. Subsymbolische Modelle haben keine natürliche Metapher, die für die Visualisierung genutzt werden könnte. In diesem Fall wurde der Raum aller Modellparameter des analytischen Verfahrens wie ein Merkmalsraum von Daten interpretiert. Prinzipiell können auf diese Weise Visualisierungstechniken für Modelle genutzt werden, die eigentlich für die Darstellung von Daten entwickelt wurden.
Die Darstellung eines Modells soll allgemein einen Bezug herstellen zwischen den Ergebnissen des Analyseprozesses und der Fragestellung, die durch diesen Analyseprozess beantwortet werden soll. Im Detail betrachtet hängen die Ziele insbesondere davon ab, welche Rolle das Modell innerhalb der Analyse innehat, und wer der Adressat dieser Darstellung ist. Dabei können Modelle danach unterschieden werden, wie direkt ihr Bezug zur Fragestellung ist, die die Analyse überhaupt begründet.
Ein Modell kann das Ergebnis einer Analyse repräsentieren, d.h. beispielsweise Begriffe definieren oder Entscheidungskriterien für den Anwendungsfall beschreiben. Unter einem umfassenden Blickwinkel stellt sich eine Analyse jedoch als Verknüpfung aufeinander aufbauender Fragestellungen dar. Ausdrücklich sollen hier daher auch solche Modelle betrachtet werden, die als Zwischenergebnis Teil der Analyse sind und daher nicht notwendigerweise an externe Entscheidungsträger vermittelt werden müssen. Dazu können auch jene Modelle gehören, aus denen keine Aussagen über die Daten folgen, sondern Aussagen über die Wahl der Verfahren und Parameter im nächsten Iterationsschritt. Ein Beispiel dafür etwa ist das Modell aus Abschnitt 3.1.4.4, mit dem die Strategie für die Wahl der Attribute des Merkmalsraums umgesetzt wird.
Durch die Darstellung eines Modells können beispielsweise folgende Ziele verfolgt werden:

- Vermittlung der Ergebnisse der Analyse an Entscheidungsträger (d.h. „telling the story“)
- Bewertung des Modells (Blanchard, [BPKG07])
- Entscheidung über die nächste Iteration im Analyseprozess
- Vergleich von Mustern im Kontext des Modells (Langton, [LPWH06])
- Direkte Manipulation des Modells (Ankerst, [AEK00])
- Manuelle Konstruktion eines Modells aus einer Hypothese (May, [MK08a])
- Identifizierung von Mustern innerhalb des Modells (d.h. Suche nach „Mustern in Mustern“, Langton, [LPWH06])

Das erste Ziel unterscheidet sich dabei in einer Hinsicht von den anderen. Der Addressat der Darstellung ist nicht der Analyst, sondern der Entscheidungsträger. In diesem Sinne handelt es sich bei einer solchen Darstellung eher um ein Medium der Kommunikation als um ein Medium der Interaktion. Burkhard [Bur04] identifiziert an diesen verschiedenen Anforderungen den Unterschied zwischen Informationsvisualisierung und Wissensvisualisierung (siehe auch Abschnitt 2.4).
Dieses Konzept gibt die Aufgabe, die nach der Modellierung unterstützt werden soll, nicht vor. Die im folgenden Kapitel umgesetzte Visualisierung eines Entscheidungsbaums unterstützt die Bewertung des Modells, die direkte Manipulation und die manuelle Modellierung einer Hypothese für die konfirmative Analyse.

3.1.6 Zusammenfassung der Systematik

Abbildung 3.13 zeigt die drei Varianten der Beeinflussung eines Analyseverfahrens durch eine Visualisierung. Die Varianten unterscheiden sich durch die Ansatzpunkte, die aus der Visualisierung heraus gesteuert werden. Durch eine Visualisierung können Verfahrensparameter (oben), Modellparameter (links) oder auch die Eingabedaten (rechts) für das automatische Verfahren verändert werden.
Die Separation zwischen Mustererkennung und Musterbeschreibung wird umgesetzt durch die Kopplung, bei der die Interaktion in der Visualisierung als Eingabe interpretiert wird. Diese kann für die „klassischen“ Aufgaben der Analyse wie Klassifikation, Clustering oder Ausreißeranalyse genutzt werden.
Eine Kopplung, in der die Visualisierung als Verfahrensparameter eines automatischen Verfahrens interpretiert werden, dient der Exponierung jener Schritte, die die Data-Mining Verfahren vorbereiten. Mit dem Anspruch, dass die Parameter einer Analyse genauso belastbar sein müssen wie ihre Ergebnisse, verbietet sich die Möglichkeit, die Analyse von willkürlich gewählten Verfahrensparametern abhängig zu machen: Dies gilt ebenso für die Wahl der Attribute, Distanz- und Gütemaße und andere Parameter.
Zuletzt sind auch Kopplungsvarianten denkbar, in denen die Modellparameter direkt interaktiv definiert werden. Dies kann sowohl in der Visualisierung der Daten geschehen, als auch innerhalb einer Visualisierung des Modells.

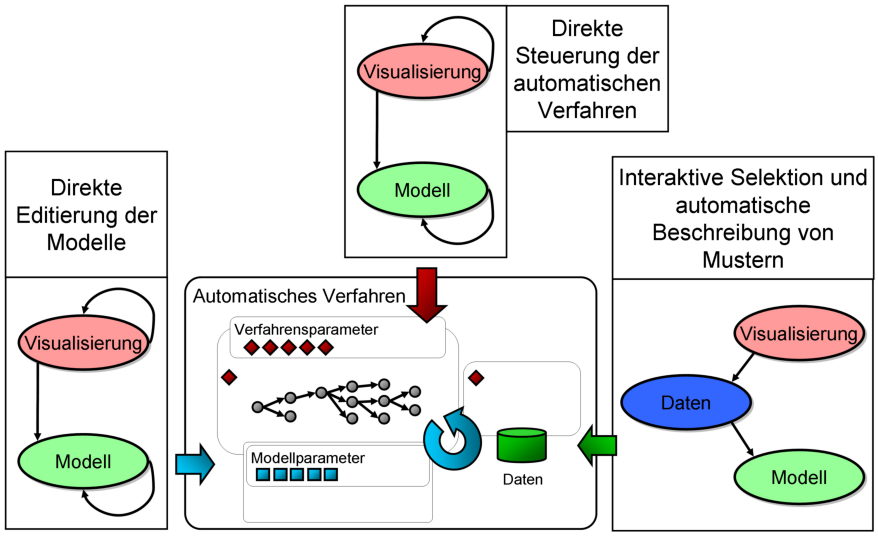


Abbildung 3.13: *Dieses Bild gibt eine Übersicht über die drei Möglichkeiten der direkten Kopplung visuell-interaktiver Verfahren mit automatischen Verfahren. Die Möglichkeiten werden unterschieden nach der Art der Ansatzpunkte des automatischen Verfahrens.*

Die Steuerung des automatischen Verfahrens durch die Visualisierung geschieht in allen drei Fällen vor dem Hintergrund der Aufgabenteilung zwischen Mensch und Maschine. Die Visualisierung reduziert dabei für den Menschen den kognitiven Aufwand. Das gilt allgemein dort, wo potentiell komplexe Bezüge zwischen Informationen hergestellt werden müssen. Das ist erstens notwendig für die Identifikation von Mustern in den Daten und zweitens für die Bestimmung von Verfahrensparametern der Analyse auf der Basis mehrerer Kriterien.
Das soll jedoch nicht bedeuten, dass der Mensch alle Aufgaben der Analyse nur mit Mustererkennung lösen könnte. Dennoch wurde hier das Repertoire der untersuchten Operationen, mit denen der Mensch zur Analyse beiträgt, auf die Mustererkennung und den Vergleich von Bildern eingeschränkt. Diese Beschränkung wird bewusst in Kauf genommen: In Abschnitt 3.1.2 wurde dargelegt, dass eine Kopplung auch zwischen sehr verschiedenen Verfahren und Werkzeugen praktisch immer möglich ist, wenn der Mensch die Interpretations- und Evaluationsleistung bei der Interaktion selbst erbringt. Durch die Interpretation werden die Informationen aus dem Kontext der Verfahren gelöst und im Kontext der Aufgabe beschrieben.
Der Nachteil dabei ist, dass die Komplexität der Interpretation ohne Hilfsmittel durch Arbeitsgedächtnis und Aufmerksamkeit beschränkt ist. Das hier vorgestellte Konzept beschreibt in erster Linie ein solches Hilfsmittel. Durch die Beschränkung soll die Suche und Identifikation jener Lösungen für die Kopplung zwischen Visualisierung und Data-Mining-Verfahren erzwungen werden, in denen der Mensch die Korrespondenz zwischen wahrgenommenem Muster und der Interpretation dieses Musters nicht selbst herstellen muss. Erst die Beschränkung erlaubt die Beantwortung der Frage, welche Aufgabenstellungen in der Analyse auch unter dieser Einschränkung noch gelöst werden können.

Die Erkennung von Mustern und Strukturen ist eine wichtige, elementare Aufgabe innerhalb der Analyse. Sie ist stets einer bestimmten Aufgabe untergeordnet. Dieses Konzept schreibt

nicht vor, wie die übergeordneten Aufgaben (z.B. „Herstellung von Bezügen“ oder „Test von Hypothesen“, siehe Abschnitt 2.4.1) in die Gesamtstrategie der Analyse eingebettet werden. Der Fokus dieses Konzepts besteht in der Vereinfachung der Mustererkennung und Beschreibung durch die Mensch-Machine-Kopplung und die Erhöhung der Sicherheit durch die Kopplung von explorativer und konfirmativer Analyse (siehe folgender Abschnitt).
In der hier vorgestellten Systematik wurden Aufgabenstellungen des Data-Mining untersucht. Diese Systematik dient auch der Einordnung eigener Verfahren, die im folgenden Kapitel im Detail beschrieben werden. Ebenso eingeordnet wurden bereits existierende Verfahren aus der Literatur; einerseits um zu zeigen, dass das Konzept tragfähig ist, und keine exotische Problemstellung beschreibt, anderseits um zu zeigen, an welchen Stellen noch technologische Lücken existieren.

Steuerung von Verfahrensparametern: Weil die Analyse iterativ abläuft, ergibt sich eine starke Wechselwirkung zwischen der eigentlichen Analyse und allen Prozessen, mit denen sie vorbereitet wird. Dabei können für die Schritte, die das Data-Mining vorbereiten, unter Umständen ebenso mächtige Verfahren erforderlich sein, wie für das Data-Mining selbst (siehe z.B. Yang und Honavar [YH98]). Wenn die Verfahrensparameter für die vorbereitenden Schritte jedoch nicht bekannt sind, erfordert dies eigentlich eine Analyse eigenen Rechts, um die Wechselwirkungen zwischen Verfahrensparametern und Ergebnissen zu identifizieren.
In Abschnitt 3.1.4.4 wurde eine Strategie für die Wahl der Attribute vorgestellt, in deren Zentrum ein Netzwerk steht, das nur die wesentlichen Informationen über die Attribute, die für diese Wahl relevant sind, zusammenfasst und abstrahiert. Die Visualisierung dieses Netzwerks soll eine Externalisierung des Modells sein, das der Anwender während der Analyse sonst in seinem Kopf (oder auf einem Blatt Papier) konstruieren würde.
Die vorgestellte Strategie für die Wahl der Attribute unterscheidet sind hinsichtlich der Herangehensweise, nach der eine mögliche Kopplung identifiziert wird, bei der eine Visualisierung als Datenquelle für ein automatisches Verfahren genutzt werden kann. Die beiden Herangehensweisen lassen sich etwa wie folgt charakterisieren:
Setzt man bei einer gegebenen Visualisierung der Daten an, besteht die Möglichkeit zu untersuchen, welche Art von Information durch die direkte Interaktion definiert werden kann, und danach zu entscheiden, mit welchen automatischen Verfahren diese Informationen weiterverarbeitet werden können und sollen. Die so entstehenden Verbindungen sind daher *Ausgabe-getrieben*. Umgekehrt kann man auch die automatischen Verfahren untersuchen, um zu bestimmen, welche Art von Information für die Konfiguration und Steuerung des Verfahrens benötigt werden, und entsprechend zu entscheiden, auf welche Weise diese Informationen dargestellt werden können. In Analogie sind die so entstehenden Verbindungen *Eingabe-getrieben*.
Kopplungen, über die eine Menge von Datensätzen des Merkmalsraums oder Modellparameter ausgetauscht werden, können stets aus beiden Perspektiven betrachtet werden. Die Unterscheidung wird dagegen dann relevant, wenn durch die Interaktion Verfahrensparameter des automatischen Verfahrens definiert werden sollen, denn diese muss immer Eingabegetrieben sein.
Verfahrensparameter haben ihrer Definition nach keine Entsprechung zu einer bestimmten Punktmenge des Merkmalsraums in dem auch die Daten liegen - sie werden also insbesondere nicht durch ein Muster der zu analysierenden Daten beschrieben. Dies bedeutet aber

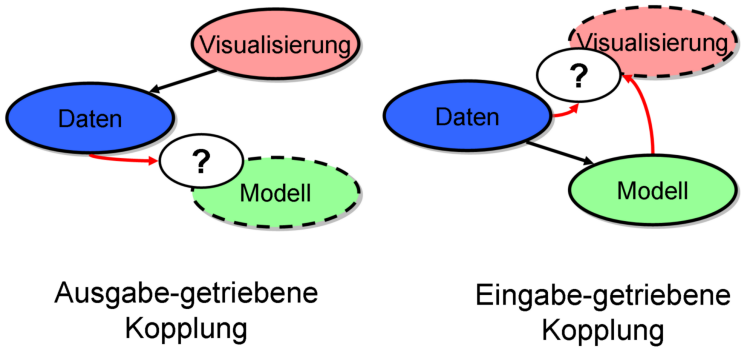


Abbildung 3.14: *Für die Etablierung einer Kopplung von visuell-interaktiven zu automatischen Techniken kann man sich einerseits an der Ausgabe der Visualisierung orientieren und automatische Verfahren identifizieren, die durch diese Daten gesteuert werden können. Umgekehrt ist es jedoch auch möglich, sich am automatischen Verfahren zu orientieren, um passende Visualisierungstechniken identifizieren.*

nicht, dass diese Parameter nicht durch direkte Interaktion in einer Visualisierung gesteuert werden könnte, sondern nur, dass dies keine Visualisierung der Daten sein muss.
Die Visualisierung des Netzwerks für die Wahl der Attribute ist ein Beispiel dafür: Eine beliebige Teilmenge von Attributen des Merkmalsraums läßt sich nicht durch ein Muster dieses Merkmalsraums beschreiben[3]. Die Fragen, die für eine Eingabe-getriebene Kopplung für die Steuerung automatischer Verfahren beantwortet werden müssen, sind zunächst folgende:

1. Was ist die zu beantwortende Fragestellung?
2. Existiert ein Bezug zwischen den vorhandenen Informationen und der zu beantwortenden Fragestellung?
3. Läßt sich dieser Bezug so darstellen, dass die Beantwortung der Frage sich aus der Darstellung ergibt oder durch die Darstellung vereinfacht wird?

Es handelt sich hierbei um die gleichen Fragen, die immer auch für die visuelle Datenanalyse insgesamt gestellt werden müssen. Es gibt daher keinen Grund, für die Beantwortung der Fragen, die *während* der Analyse erst auftauchen, nicht das gleiche Repertoire an Techniken einzusetzen, wie für die Fragestellung, die eine Analyse letztbegründet - d.h. die Problemstellung oder Fragestellung des Anwenders.
Im Beispiel wurde eine Netzwerkvisualisierung gewählt, die einen Kontext herstellt, der dem Abstraktionsniveau der Fragestellung entspricht. Die Frage, welche Attribute als Nächstes für welches Verfahren eingesetzt werden sollen, wird umgesetzt auf ein Problem der Netzwerkanalyse.
Motivation für die Unterscheidung zwischen Ausgabe-getriebenen und Eingabe-getriebenen Strategien für die Steuerung von automatischen Verfahren ist die Darstellung der Möglichkeiten, die sich aus dem Wechsel der Perspektive ergeben. Die Wahrnehmung der Steuerung von

[3]Ein erkanntes Muster kann natürlich sehr wohl dazu verwendet werden, um *automatisch* eine Menge von Attributen aus dem Merkmalsraum zu extrahieren. Gemeint ist hier aber nicht die *Ausgabe*, die ein automatisches Verfahren berechnen könnte, sondern die *Eingabe*, über die es gesteuert wird. Vor der Anwendung von Data-Mining Verfahren müssen die Eingabedaten aus einer Visualisierung herleitbar sein.

automatischen Verfahren als Problemstellung eigenen Rechts ermöglicht die Auseinandersetzung mit der Steuerung unter den gleichen Prämissen, die auch für die Analyse insgesamt gelten. Die Untersuchung von Verfahrensparametern innerhalb des Kontexts, in dem sie entschieden werden, macht die Metaanalyse aus.
Die zentrale Annahme für die Metaanalyse besteht darin, dass ein Anwender in der Lage sein muss, den Effekt bestimmter Verfahren und Verfahrensparameter zu vergleichen. Wenn das Ergebnis einer Analyse bei gleichen Daten instabil ist bezüglich der Wahl von Verfahren und Parametern, dann ist das Ergebnis mindestens auch ein Artefakt dieser Wahl. Überspannt eine Analyse mehrere Iterationen und einen längeren Zeitraum, kann der Bezug zwischen Entscheidung und Effekt nicht allein durch den Menschen hergestellt werden. Die Annahme motiviert die Suche nach einer Darstellung von Verfahrensparametern, in der sich diese Bezüge als Muster manifestieren können.

Veränderung des Suchraums: Betrachtet man die Aufgabenteilung zwischen Mensch und Maschine aus der Sicht der automatischen Verfahren wird durch die Interaktion des Menschen der Suchraum für die Heuristiken verändert. Das Ziel einer Veränderung des Suchraums besteht immer darin, die Wahrscheinlichkeit dafür zu erhöhen, dass die Heuristik für das Modell das globale Optimum bezüglich der Gütefunktion identifiziert.
Im letzten Kapitel (siehe Abschnitt 2.3.4) wurden bereits zwei Strategien identifiziert, mit denen der Suchraum der Heuristiken verkleinert werden kann:

- Externe Bestimmung von Modellparametern
- Initialisierung von Modellparametern

Im ersten Fall wird der Suchraum explizit verkleinert, wodurch der Heuristik dieser Freiheitsgrad für die Optimierung entzogen wird. Im zweiten Fall wird der Suchraum implizit verkleinert. Eine Initialisierung ist nur dann sinnvoll, wenn es sich bei den entsprechenden Modellparameter um einen numerischen Parameter handelt. Der Nutzen einer Initialisierung gründet darauf, dass dann ein Modellparameter bereits „hinreichend nahe“ am globalen Optimum liegt. Dazu muss für die Heuristik jedoch ersichtlich sein, wodurch sich „hinreichend nahe“ Punkte des Parameterraums von „entfernten“ Punkten unterscheiden. Der Suchraum wird unter der Annahme verkleinert, dass die Initialisierung innerhalb des Konvergenzradius um das globale Optimum stattfindet.
Für beide Strategien können Beispiele benannt werden, in denen sie durch die Interaktion in einer Visualisierung umgesetzt werden. Eine externe Bestimmung von Modellparametern wird beispielsweise bei Ankerst et al. [AEEK99] für die Konstruktion von Entscheidungsbäumen durchgeführt. Die Definition von Unterteilungspunkten für die Partitionierung der Attribute eines Knotens ersetzt den Teil der Heuristik, in dem gerade die Unterteilungspunkte bestimmt werden. Dies kann im Extremfall gerade so weit gehen, dass sämtliche Freiheitsgrade eliminiert werden, und die automatische Heuristik mithin obsolet wird (siehe Liu [LSG04]).
Eine Initialisierung von Modellparametern kann beispielsweise bei der interaktiven Bestimmung der Startwerte für ein Clusteringverfahren erfolgen (siehe z.B. [CL04]). Eine andere Initialisierung beschreiben wiederum Ankerst et al. [AEK00], wo die ersten Knoten des

Entscheidungsbaums durch den Anwender definiert werden, die Verfeinerung der Modells jedoch über automatische Verfahren geschieht. Durch die Interpretation der Interaktion als

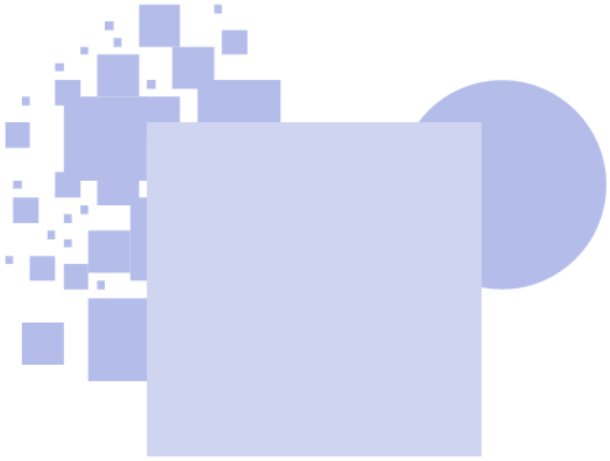


Abbildung 3.15: *Bei der Wahrnehmung von Mustern wird der Bildbereich segmentiert. Für ein automatisches Verfahren kann dies als Vorverarbeitung betrachtet werden mit dem Ziel, möglichst einfache Strukturen zu bestimmen. Diese Strukturen lassen sich wiederum in einfache Gütefunktionen umsetzen. Die Abgrenzung verschiedener Strukturen (rechts) und die Abgrenzung von Rauschen (links) verhindert, dass bei der Optimierung „konkurrierende" Strukturen einen Kompromiss bei der Musterbeschreibung erfordern. Dies kann insbesondere dann genutzt werden, wenn die Beschreibung der einzelnen Strukturen jede für sich genommen einfacher ist, als die gemeinsame Beschreibung aller Strukturen.*

Eingabedaten eröffnet sich eine dritte Strategie für die Veränderung des Suchraums. Die Eingabedaten definieren die Gütefunktion für die Heuristik. Schöneburg et al. [SHF96] vergleichen die Performance von Heuristiken (in diesem Fall von Genetischen Algorithmen bzw. Evolutionsstrategien) angewendet auf verschiedene numerische - allesamt stetige - Testfunktionen. Dabei zeigt sich, dass die Eigenschaften der Gütefunktion - wie etwa die Anzahl der lokalen Optima und derer Konvergenzradien - einen großen Einfluss auf die Ergebnisse der Heuristiken haben.
Bei der Mustererkennung ist der Mensch in der Lage, Bildbereiche voneinander zu separieren. Muster können dabei vom Rauschen und voneinander abgegrenzt werden, wobei die Abgrenzung so erfolgt, dass möglichst einfache Strukturen entstehen. Der Mensch kann dabei gut seine Aufmerksamkeit auf ein Muster fokussieren. Eine solche Fokussierung ist bei automatischen Verfahren nicht möglich. Sobald ein Datensatz mehrere Strukturen enthält, die sich nicht durch das gleiche Modell optimal beschreiben lassen, können Gütefunktion und Heuristik jeweils nur eine Kompromisslösung finden, die die Beschreibungsfehler „im Mittel" reduziert.
Dieses Problem wird zusätzlich dadurch verschärft, dass unterschiedliche Strukturen des Datensatzes nicht durch das gleiche Verfahren „gut" beschrieben werden können. In diesem Fall stehen die Strukturen des Datensatzes in Konkurrenz hinsichtlich der Komplexität des Modells, das die Strukturen beschreiben soll.
Ziel der Mustererkennung ist - aus der Sicht automatischer Verfahren - eine Vorsegmentierung der Daten. Auf diese Weise repräsentiert die Gütefunktion wenige, möglichst einfache Strukturen, so dass mindestens die Fälle, in denen die Optimierung lediglich einen Kompro-

miss herstellt, vermieden werden können. Dies allein garantiert nicht, dass die Beschreibung auch außerhalb des Kontexts des gewählten Verfahrens optimal ist. Um die Qualität des Verfahrens direkt bewerten zu können, vervollständigt das iterative Feedback das hier vorgestellte Konzept. Dieses wird im folgenden Abschnitt beschrieben.

3.2 Konfirmatives Feedback

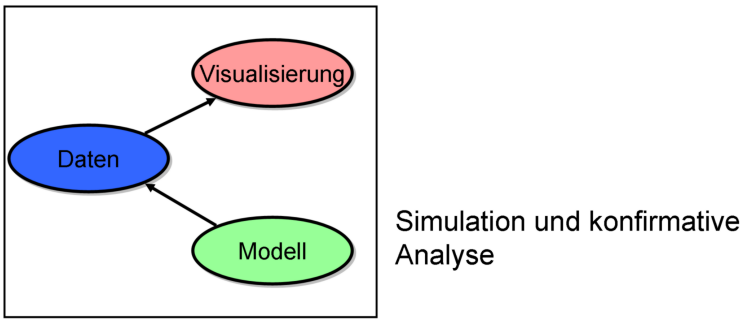


Abbildung 3.16: *Diese Erweiterung beschreibt die Rückkopplung, bei der ein erzeugtes oder vorgegebenes Modell verwendet wird, um durch Simulation einen synthetischen Datensatz zu erzeugen. Eine Methode der konfirmativen Analyse ist ein visueller Abgleich zwischen erzeugten Daten und Referenzdaten, in dem qualitative Fehler bei der Modellierung identifiziert werden können.*

Im letzten Kapitel (siehe Abschnitt 2.3.2) wurde dargelegt, dass eines der zentralen Probleme bei der automatischen Modellgenerierung darin besteht, dass jedes automatische Verfahren stets nur die Muster identifizieren kann, die innerhalb der Modellfamilie hinreichend einfach beschrieben werden können. Innerhalb des Modells kann die Güte dieser Beschreibung nur quantitativ bestimmt werden - etwa durch den Prädiktionsfehler, der bei der Anwendung des Modells auf Testdaten bestimmt werden kann. Ein systematischer Modellierungsfehler kann innerhalb der Modellfamilie jedoch nicht identifiziert werden.
Die Trennung zwischen Mustererkennung und Musterbeschreibung löst dieses Problem jedoch nur zur Hälfte: Ein menschlicher Nutzer ist nicht determiniert, was die Klasse der erkennbaren Muster angeht. Dass dieses Muster geeignet beschrieben wurde, kann jedoch allein *durch die Beschreibung selbst* nicht bewiesen werden. Die Qualität der Beschreibung kann sich insofern verbessern, da zum Beispiel die Trennung zwischen Mustern und Rauschen durch den Menschen präziser durchgeführt wird, als durch jede Maschine. Fehler bei der Wahl der Modellfamilie bleiben davon jedoch unbeeinflusst und müssen durch konfirmative Analyse belastet werden.
Die Strategie für die konfirmative Analyse ist nicht für alle Verfahren gleich. Im Wesentlichen müssen prädiktiven und deskriptiven Verfahren unterschiedlich gehandhabt werden. Breiman [Bre01b] betont in seiner Diskussion, dass das offensichtlichste Qualitätsmerkmal für prädiktive Verfahren auf der Genauigkeit der Prädiktion beruhen muss, wobei die Genauigkeit gegen die Modellkomplexität abgewogen werden muss. Bei deskriptiven Verfahren ist dies in dieser Form nicht möglich. Die grundsätzliche Strategie für das konfirmative Feedback besteht jedoch in beiden Fällen darin, das beschriebene Modell mit dem Muster zur Deckung zur bringen, das über die Interaktion der Auslöser für die Analyse war.
Das Ziel ist hier, den Abgleich zwischen wahrgenommenem Muster und beschriebenen Modell innerhalb einer Visualisierung so zu gestalten, dass er sich auf den Vergleich zweier Bilder reduziert. Der Vergleich von Hypothesen oder Modellen erfordert im Normalfall kognitiven Aufwand. Ebenso wie bei der subsymbolischen Interaktion für die explorative Analyse, soll hier untersucht werden, wie dieser kognitive Aufwand reduziert werden kann.

3.2.1 Konfirmative visuelle Analyse für prädiktive Modelle

Prädiktive Verfahren liefern ein Modell, das eine Funktion $\Phi : X \rightarrow Y$ beschreibt, welche Tupel unabhängiger Attribute X eines Merkmalsraums M auf abhängige Attribute Y abbildet. Die Funktion beschreibt einen entsprechenden Zusammenhang innerhalb des Merkmalsraums. Dabei ist es im Prinzip unerheblich, ob das Modell tatsächlich das Ergebnis eines Data-Mining Verfahren ist, oder ob es aus einer Hypothese abgeleitet wurde, die diesen Zusammenhang explizit konstatiert.
Das Modell kann angewendet werden auf jeden Datensatz dessen unabhängige Attribute $x \in X$ bekannt sind. Das Modell kann geprüft werden gegen jeden Datensatz, für den zusätzlich auch die abhängigen Attribute $y \in Y$ bekannt sind und eine Referenz für den Vergleich darstellen können. Beim visuellen Feedback ist jedoch der Fehler bezüglich eines einzelnen Datensatzes irrelevant, weil dieser keine Rückschlüsse auf einen systematischen Modellierungsfehler erlaubt.
Das Konzept für das visuelle Feedback für prädiktive Verfahren beruht auf dem Prinzip, die Abweichungen zwischen $\Phi(x)$ und y ebenfalls als Muster darzustellen. Handelt es bei den betrachteten Attributen ausschließlich um numerische Verfahren kann die als visuelle *Analyse der Residuen* verstanden werden. Da die Prädiktion $\Phi(x)$ und die Referenz y den gleichen Datentyp besitzen, kann durch jede Visualisierungstechnik, mit dem die abhängigen Attribute Y dargestellt werden können, auch die Prädiktion $\Phi(X)$ dargestellt werden. Entscheidend für das visuelle Feedback ist, dass man beide Darstellungen gegenüberstellt.
Der Vergleich des Modells mit dem Referenzdaten reduziert sich auf den Vergleich zweier Bilder. In vielen Fällen ist es sogar möglich, beide Darstellungen zu überlagern, so dass sich die Differenzen zwischen Modell und Referenzdaten entweder als Muster (im Fall eines systematischen Fehlers), oder als Rauschen (im Falle irrelevanter Schwankungen) manifestieren können.
Zwei Bedingungen muss man bei der Darstellung berücksichtigen. Die erste Bedingung betrifft die Tatsache, dass Y nicht notwendigerweise numerisch sein muß. Das ist beispielsweise bei Klassifikationsverfahren der Fall. Eine Differenz $(y - \phi(x))$ kann dann, im Gegensatz zu Regressionsverfahren, nicht bestimmt werden. Anstatt die Werte (und etwaige Differenzen) direkt darzustellen, kann man beispielsweise aber auch die Verteilung der Werte von Y miteinander vergleichen, wobei stets die Daten mehrerer Datenelemente zusammengefasst werden müssen. Ein Beispiel dafür illustriert Abbildung 3.17.
Die zweite Bedingung ergibt sich dadurch, dass zwei Bilder nur dann verglichen werden können, wenn der Mensch eine Korrespondenz zwischen diesen Bildern herstellen kann. Wenn die Daten, die in beiden Bildern miteinander verglichen werden sollen, erst gesucht werden müssen, dann würde der Abgleich Aufmerksamkeit und damit kognitive Ressourcen erfordern.
Die jeweils korrespondierenden Elemente haben die gleichen unabhängigen Attribute X. Daher sollten diese Attribute auch auf die visuellen Attribute abgebildet werden, durch die die Korrespondenz visuell hergestellt wird. Oft, aber nicht grundsätzlich, ist das die Position. Die Werte Y bzw. $\Phi(X)$ sollten dementsprechend nicht auf die Positionen abgebildet werden[4]. Dieses einfache Prinzip ist die Grundlage für die konfirmative visuelle Analyse. Im

[4]Ein Gegenbeispiel wäre ein Liniengraph, der die Prognose eines zeitlichen Verlaufs darstellt. Hier definiert die Abszisse eine Korrespondenz zwischen verschiedenen Positionen, die direkt miteinander verglichen werden

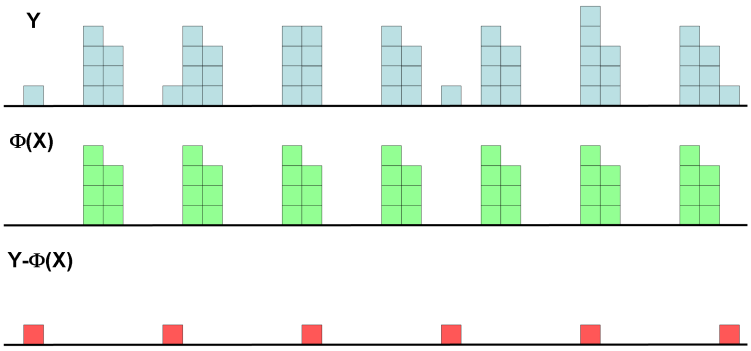


Abbildung 3.17: *Wie explorative und konfirmative Analyse ineinander greifen, zeigt sich in diesem Beispiel. Dabei beschreibt Y die Originaldaten. $\Phi(X)$ bezeichnet die Verteilung nach dem prädiktiven Modell. Jedes Kästchen steht hier für einen Datensatz, verteilt nach dem Wertebereich eines Attributs. Im Prinzip ist es unerheblich, ob Φ durch ein Data-Mining-Verfahren oder durch eine Hypothese formuliert wurde. Während die statistischen Fehlermaße* (Precision, Recall, etc.) *nur quantitative Gütemaße liefern, liefert der visuelle Vergleich von Modell und Daten qualitative Informationen über das Modell. In diesem Fall handelt es sich um einen systematischen Fehler, der auch eine Indikation für die Verfeinerung des Modells liefert. Die Suche nach Mustern (= explorative Analyse) beim Abgleich von Modellen und Daten (=konfirmative Analyse) verbindet diese beiden zentralen Strategien der Analyse. Bei ausschließlich numerischen Daten ist das die Analyse der Residuen. Prinzipiell möglich ist dies jedoch auch bei nicht-numerischen Daten, wenn man anstelle der Differenzen der Werte, die Differenzen der Werteverteilung analysiert.*

Unterschied zur konfirmativen statistischen Analyse wird der Prädiktionsfehler nicht durch statistische Kennzahlen beschrieben, die in erster Linie quantitative, jedoch keine qualitativen Aussagen über systematische Fehler erlauben. Problematisch ist zudem, dass die Messung des Prädiktionsfehlers - insbesondere bei der Regression - auf Differenzfunktionen beruht (siehe 3.1.4.6), deren Bezug zu den Daten und deren Einfluss auf das Ergebnis belastbar untersucht werden muss.
Durch die Beschreibung des Prädiktionsfehlers über eine oder wenige statistische Kennzahlen werden die Informationen über die Qualität der Prädiktion in eine möglichst kompakte Form aggregiert. Dies stellt jedoch das Ende eines weiten Spektrums dar. Durch die hohe Aggregation gehen charakteristische Eigenschaften des Fehlers verloren, die für eine Bewertung des Modells und ggf. eine Verfeinerung des Prädiktors verwendet werden können. Hinter der Suche nach systematischen Fehlern steht dagegen die Frage, ob Eigenschaften existieren, nach denen die Datensätze, für die das Modell korrekte Ergebnisse liefert, von den Datensätzen unterschieden werden können, für die es falsche Ergebnisse liefert. Beim Hinterfragen der Kriterien für die Qualität eines Modells werden damit die Fragestellungen der konfirmativen und der explorativen Datenanalyse eng miteinander gekoppelt.

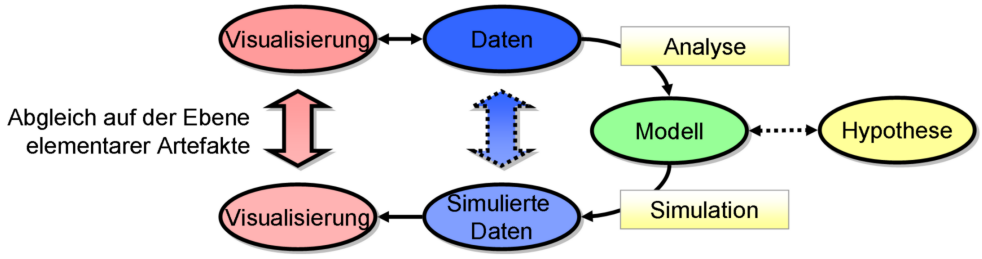


Abbildung 3.18: *Das visuelle Feedback beruht auf den zwei komplementären Prozessen* Modellkonstruktion *(d.h. die Anwendung eines Data-Mining-Verfahrens) und* Simulation. *Diese beschreiben jeweils eine Transformation von elementaren Artefakten hin zu Modellen und zurück. Bei einem getreuen prädiktiven Modell kann man erwarten, dass Beschreibungsfehler eliminiert werden. Durch das Feedback kann der Modellierungsprozess überwacht werden. Wenn die beiden Prozesse hintereinander ausgeführt nicht die Muster reproduzieren, die der Anwender in der Visualisierung der Referenzdaten erkennt, müssen das Verfahren für die Analyse oder dessen Einstellungen verworfen werden.*

3.2.2 Nutzung von Techniken der explorativen Analyse für die konfirmative Analyse

Der Vorteil der konfirmativen visuellen Analyse ist, dass ein visueller Kontext geschaffen wird, innerhalb dessen die Ergebnisse des Prädiktors und die Referenzdaten untersucht werden können. Zu berücksichtigen ist dabei, dass *jedes* Attribut des Merkmalsraums, das einem Datensatz eindeutig zugeordnet werden kann, eine potentielle Eigenschaft darstellt, anhand derer die Qualität des Prädiktors bewertet werden kann. Potentiell relevant sind also die abhängigen Attribute Y, die unabhängigen Attribute X, aber auch alle anderen Attribute des Merkmalsraums M. Letzteres ist insbesondere dann der Fall, wenn bei der Konstruktion des Modells a-priori bestimmte Attribute in den Daten unberücksichtigt blieben.
Die konfirmative Analyse eines prädiktiven Modells Φ kann in der Form einer speziellen explorativen Analyse umformuliert werden. Das Ziel der Analyse ist jedoch nicht die Suche nach Beziehungen zwischen Attributen im Merkmalsraum. Vielmehr soll die Sicherheit dafür erhöht werden, dass *keine* relevante Beziehung zwischen dem Merkmalsraum M und den Differenzen $dist(\Phi(x_i), y_i)$ für die Datensätze $(x_i, y_i)_i$ existiert. Dies zu beweisen, ist nur durch erschöpfende Suche möglich - durch die Untersuchung möglichst vieler potentieller Beziehungen kann die Unsicherheit nur verkleinert, jedoch niemals ganz eliminiert werden.
Aus dieser Umformulierung ergibt sich der Vorteil, dass alle Techniken, die für explorative Analyse einsetzbar sind, auch auf diese Aufgabe angewendet werden können. Allerdings muss man berücksichtigen, dass man damit gleichsam auch die Herausforderungen der explorativen Datenanalyse übernimmt. Automatische Verfahren sind im Normalfall nicht beschränkt hinsichtlich der Anzahl der Attribute, die sie verarbeiten können. Allerdings kann mit automatischen Verfahren höchstens nachgewiesen werden, dass eine bestimmte Klasse von Mustern nicht im Datensatz vorkommt - im Allgemeinen mit einem Unsicherheitsfaktor. Bei visuell-interaktiven Techniken ist es gerade umgekehrt. Diese sind beschränkt in der Anzahl der Attribute, die gleichzeitig dargestellt werden können, jedoch kann ein Mensch potentiell

können - nämlich jene, die untereinander liegen. Die Werte des Prädiktors und die Originaldaten können dann auf der Ordinate abgetragen werden.

mehr Muster identifizieren.
Eine neue Methodik vorzuschlagen, die sich mit diesem speziellen Fall der explorativen Datenanalyse auseinandersetzt erscheint nicht sinnvoll. Es ist anzunehmen, dass jede Technik und jede Methode, mit der die Suche nach Mustern verbessert werden kann, auch die Nachweissicherheit erhöht, dass in einem Datensatz *keine* relevanten Muster enthalten sind. Das schließt automatische, visuell-interaktive Techniken und deren Kopplung ein.

3.2.3 Nutzung von Techniken der konfirmativen Analyse für die explorative Analyse - visuelles Feedback

Die Unterstützung und Verfeinerung der konfirmativen Analyse durch explorative Analyse funktioniert auch in die andere Richtung: Das visuelle Feedback nutzt Methoden der konfirmativen visuellen Analyse, um die Musterbeschreibung zurückzuführen auf die Mustererkennung und kehrt damit den Prozess aus dem ersten Teil des Konzepts um.
Die Motivation für das visuelle Feedback ist die Bestätigung, dass mit dem automatischen Verfahren genau das Muster beschrieben wird, das der Mensch innerhalb der Visualisierung erkannt und durch die Interaktion beschrieben hat. Die Präsentation der Musterbeschreibung ermöglicht zwar ebenfalls eine Bestätigung, allerdings müsste der Mensch dabei die Korrespondenz zwischen Modell und Muster selbst herstellen. Wie in den vorigen Abschnitten beschrieben, ist es jedoch vergleichsweise einfach, bei einem prädiktiven Modell diese Korrespondenz durch die Anwendung des Prädiktors auf die Daten herzustellen, die in der Visualisierung dargestellt werden.
Hervorhebung (*„Highlighting“*) gehört zu den wichtigsten Methoden für das Feedback auf Nutzerinteraktion im Arbeitsbereich. Dies gilt insbesondere auch beim Feedback auf die direkte Selektion innerhalb der gleichen Visualisierung. Mit dem visuellen Feedback für die explorative Analyse soll das Hervorheben auf Methoden der explorativen Analyse verallgemeinert werden. Das Hervorheben nach der direkten Selektion beruht auf der Bestimmung des Urbilds (siehe 2.4.3.1) eines Bildelements oder eines Bildausschnitts. Dabei wird das Datenobjekt oder der Wertebereich bestimmt, auf den sich die Interaktion bezieht. Danach werden diese mit modifizierten visuellen Attributen dargestellt (siehe Abbildung: 3.19).

Wenn S eine Menge von Datenobjekten des Merkmalsraums M bezeichnet und V die visuelle Abbildung vom Merkmalsraum in den Bildraum, dann definierte V(S) das dargestellte Bild. Bei der direkten Selektion kann der Anwender eine Menge $P \subset S$ von Datenobjekten nicht direkt auswählen, sondern nur deren Abbilder $V(P) \subset V(S)$. Über das Urbild der visuellen Abbildung wird die ausgewählte Teilmenge $P = V^{-1}(V(P))$ des Merkmalsraums bestimmt. Wie in Abschnitt 2.4.3.1 beschrieben, kann die Umkehrabbildung V^{-1} in manchen Visualisierungstechniken für beliebige Teilmengen $P \subset M$ des Merkmalsraums definiert werden.
P ist die Teilmenge, die das vom Nutzer wahrgenommene Muster konstituiert. Das Feedback auf direkte Selektion funktioniert dadurch, dass das Bild von P genau mit dem selektierten Bereich zur Deckung kommt.
Das Konzept für die Bestimmung des Urbilds wird hier erweitert auf ein prädiktives Modell $\Phi : M \rightarrow Y$. Dabei ist Y ein synthetisches Attribut, dass die Zugehörigkeit eines Punktes zum Muster P indiziert. Der Einfachheit halber sei zunächst angenommen, das Y stets

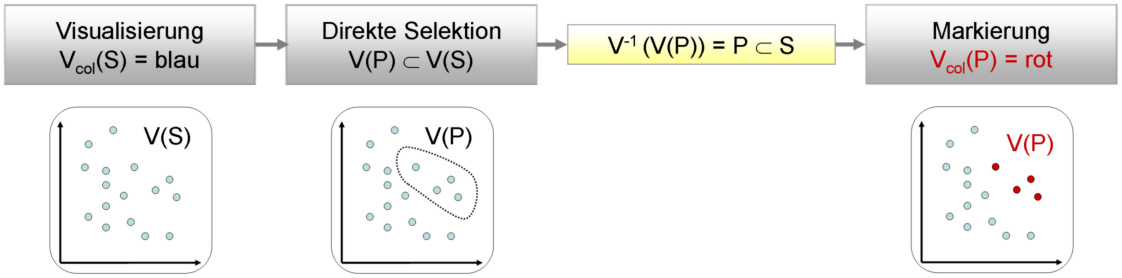


Abbildung 3.19: *Das Hervorheben von Datenobjekten oder Bereichen gehört zum am häufigsten verwendeten Feedback auf die direkte Selektion durch den Anwender. Da der Anwender die Datenobjekte nicht direkt selektiert, sondern nur deren Bilder, beruht die direkte Selektion - wie auch* Linking & Brushing *(siehe auch Abbildung 2.15) - auf der Möglichkeit, zu einer visuellen Abbildung V auch das Urbild V^{-1} zu bestimmen. Für die Markierung muss ein entsprechend geeignetes visuelles Attribut (hier: Farbe) vorgesehen werden. Beim visuellen Feedback wird die Hervorhebung auf analytische Modelle verallgemeinert. Dafür wird nicht nur das Urbild der Visualisierung V^{-1} bestimmt, sondern auch das Urbild eines analytischen Modells Φ^{-1}, das ein Muster P repräsentiert.*

Informationen beschreibt, die sich auf einziges Muster beziehen. In diesem Fall gilt daher $Y = \{pattern, nopattern\}$. Der Klassifikator soll dabei folgende Gleichungen erfüllen:

$$\Phi(x) = \begin{cases} pattern & \text{if } x \in P, \\ nopattern & \text{if } x \notin P. \end{cases} \tag{3.1}$$

Die Unterscheidung von Muster und Nicht-Muster ist eine Abstraktion, die es jedoch erlaubt, Klassifikations- und Regressionsmodelle wie auch Clusteringverfahren gleichermaßen zu beschreiben. Wenn man erlaubt, mehrere Muster zu beschreiben, erhöht sich Anzahl der Werte von Y entsprechend; das Prinzip bleibt jedoch das Gleiche.
Hier wird nicht vorgegeben, mit welchen Verfahren der Klassifikator Φ berechnet wird - das visuelle Feedback dient ja gerade der Bewertung dieses Prädiktors und des Verfahrens. Für das visuelle Feedback basiert nun auf der Bestimmung des Urbilds

$$\tilde{P} = \Phi^{-1}(pattern) = \{x \in M | \Phi(x) = pattern\} \tag{3.2}$$

Hier wird nicht vorausgesetzt, dass dieses Urbild analytisch bestimmt werden kann. Vielmehr ist die Invertierung nicht anderes als die Anwendung des Modells auf Punkte S des Merkmalsraums, wobei die Teilmenge $\tilde{P} \subset S$ bestimmt werden kann. Dieser Prozess soll hier - unabhängig davon, ob das Modell zeitliche Beziehungen beschreibt oder nicht - *Simulation* genannt werden.
$\tilde{P}$ beschreibt das Muster in der Form, wie es durch das Modell beschrieben wird. Da es sich ebenfalls um eine Teilmenge des Merkmalsraums handelt, kann diese in der gleichen Form hervorgehoben werden, wie es beim Feedback auf die direkte Selektion für das Muster P. Die Bestimmung des Prädiktors und seiner Umkehrabbildung erscheint wie eine unnötige Komplikation, allerdings muss hier deutlich betont werden, dass $\tilde{P}$ im allgemeinen *nicht* mit P identisch ist. Denn dies wäre gleichbedeutend damit, dass ein automatisches Verfahren immer alle wahrnehmbaren Muster getreu beschreibt. In Abschnitt 2.3.3 wurde bereits dargelegt, dass ein automatisches Verfahren, eine beliebige Teilmenge P durch eine Teilmenge $\tilde{P}$ approximiert, die als Modell des Verfahrens beschrieben werden kann.

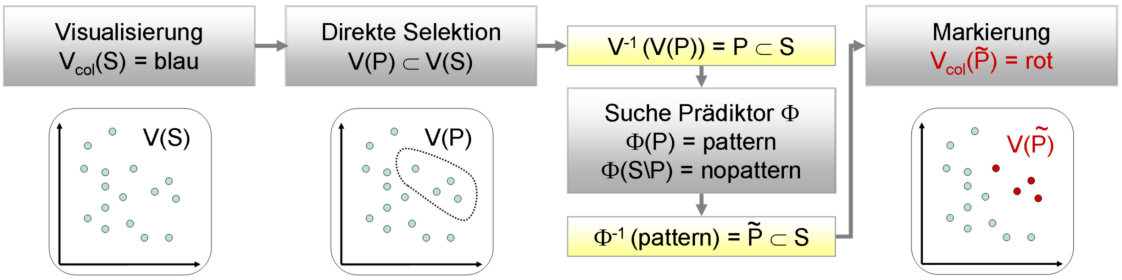


Abbildung 3.20: *Anstatt das Urbild P der visuellen Abbildung direkt zu markieren, ist es ebenso möglich, das Urbild $\tilde{P}$ zu markieren, mit dem ein Prädiktor die Menge P approximiert. Durch dieses Feedback kann der Prädiktor charakterisiert werden, denn im Allgemeinen sind P und $\tilde{P}$ nicht identisch. Der Unterschied kann als Fehler bei der Approximation durch das Modell Φ aufgefasst werden. Durch den Abgleich zwischen wahrgenommenem und reproduziertem Muster kann der Anwender untersuchen, ob dieser Modellierungsfehler nur Rauschen oder selbst ein Muster ist.*

In dieser einfachen Form kann das Feedback nicht direkt auf Verfahren angewendet werden, die Modelle konstruieren, die beliebig komplex sein können und bei denen daher die Gefahr der Überanpassung besteht. Ein Beispiel dafür wären Entscheidungsbäume mit unbeschränkter Tiefe. In diesem Fall wäre es so, dass das Modell Φ die Menge P getreu beschreibt. In diesem Fall gelte $P = \tilde{P}$ und das Feedback wäre identisch zum Feedback auf eine direkte Selektion.
Das ist deshalb problematisch, weil das Feedback in diesem Fall zwar Informationen über die Qualität des Modells, jedoch nicht über die Eigenschaften des Verfahrens liefert. Durch das Feedback sollen dem Menschen drei Dinge ermöglicht werden:

- Es soll überprüfbar sein, ob die Form des Urbilds $\tilde{P}$ mit dem wahrgenommenen Muster P zur Deckung gebracht werden kann.
- Der Anwender soll das analytische Verfahren anhand des Konvergenzverhaltens zwischen $\tilde{P}$ und P bewerten können.
- Es soll überprüfbar sein, ob und mit welchem Aufwand das Klassifikationsmodell das wahrgenommene Muster beschreiben kann.

Das erste Ziel ist durch das Feedback bereits erfüllt. Das zweite Ziel bedeutet keine Untersuchung des „Konvergenzverhaltens“ im mathematischen Sinn, denn dies würde ein Modell für eine Abstandsfunktion erfordern, sondern das Konvergenzverhalten im visuellen Sinn. Es unterscheidet sich von visuellen Abgleich darin, dass das System nicht nur ein Feedback auf ein *vollständig* definiertes, wahrgenommenes Muster liefert, sondern auf jede Interaktion reagiert, mit der der Anwender das Muster Schritt für Schritt definiert. Dieser iterative Prozess wird im folgenden Abschnitt 3.2.3.1 beschrieben.
Das dritte Ziel wird dadurch erreicht, dass die Markierung des Modells nicht nur eine binäre Unterscheidung repräsentiert, ob ein Punkt bzw. ein Datenobjekt des Merkmalsraums zum Muster gehört oder nicht. Im übernächsten Abschnitt 3.2.3.2 wird daher eine Erweiterung vorgeschlagen, in der die lokale Komplexität des Modells in das Feedback mit einfließt.

3.2.3.1 Iteratives Feedback

Im Prinzip kann der Mensch das wahrgenommene Muster P vollständig definieren und das Urbild $\tilde{P}$ des Musters im Prädiktor hervorheben lassen. Dies wurde für ein Muster nur einen Interaktions- und Feedbackzyklus erfordern, hat jedoch folgende Nachteile.

- Der Muster muss vom Anwender vollständig selektiert werden, was einen gewissen Aufwand bedeuten kann.
- Der Anwender hat während dieser Zeit kein Feedback darüber wie gut oder schlecht verschiedene (Unter-)Strukturen oder Details des Musters durch das Modell beschrieben werden können.

Beim iterativen Feedback geht man von der Annahme aus, dass ein Muster nicht einfach als eine homogene, monolithische Struktur wahrgenommen wird, sondern dass der Anwender verschiedene Unterstrukturen und verschiedene Detailgrade bei der Mustererkenung zusammenfassen kann. Je nachdem, auf welche Aspekte eines Musters der Anwender seine Aufmerksamkeit richtet, treten verschiedene Details unterschiedlich stark hervor. Man kann jedoch im Allgemeinen nicht davon ausgehen, dass alle diese Details gleich gut von einem Verfahren beschrieben werden können.
Aus diesem Grund wird hier ein iterativer Prozess vorgeschlagen, in dem auf jede Interaktion ein Feedback über den Klassifikator erfolgen soll, bei dem der Anwender und das Klassifikationsverfahren gleichzeitig zur Optimierung beitragen können. In diesem iterativen Prozess entsteht eine Folge $P_0, P_1, P_2, \ldots, P$ und eine entsprechende Folge $\tilde{P}_0, \tilde{P}_1, \tilde{P}_2, \ldots, \tilde{P}$. In diesem Prozess kann der Mensch überwachen, wie das Klassifikationsverfahren auf seine Interaktion „reagiert“, in dem sich die Grenze zwischen beschriebenem Muster und Nicht-Muster verschiebt. Dadurch gewinnt man die Möglichkeit, mit dem Fokus der Wahrnehmung auch den Fokus für die Modellierung zu steuern. Dabei können zum Beispiel nur gezielt die Teile des Modells verändert werden, die am deutlichsten vom wahrgenommenen Muster abweichen. Umgekehrt kann der Nutzer auch die Modellierung gezielt auf jene Teile des Musters fokussieren, die bereits gut modelliert werden. Dazu gehört insbesondere auch die Möglichkeit, in einer Iteration Teile des Musters wieder zu entfernen.
Der Anwender kann die Qualität des Klassifikationsmodells anhand verschiedener Ergebnisse dieses Prozesses bestimmen, die nur erkennbar sind, wenn dieser iterativ abläuft (siehe Abbildung 3.21):

- Das Feedback konvergiert in wenigen Schritten hinreichend genau gegen das wahrgenommene Muster.
- Das wahrgenommene Muster kann nur mit hohem Aufwand genau beschrieben werden (evtl. nur stellenweise in der Nähe seines Randes). Ein entsprechendes Klassifikationsmodell ist ungeeignet, wenn mit dem gleichen Aufwand auch das Rauschen im Datensatz modelliert werden könnte (siehe dazu Abbildung 2.7), und es mithin keine „natürliche“, globale Grenze für die Überanpassung gibt.

- Der Prädiktor reagiert nicht vorhersehbar auf die Eingaben des Nutzers. Instabilität des Feedbacks wäre ein Indikator für ein Modell, das gleichermaßen instabil auf Eingaben dieses Musters reagiert (und daher evtl. ebenfalls unbrauchbar sein könnte). Ein Beispiel dafür wären zu einfache Klassifikatoren, wie etwa eine einzelne Hyperebene für die Trennung einer nicht linear separierbaren Trainingsmenge.

- Durch das Feedback separiert der Anwender das Muster visuell in mehrere Teilmuster, von denen mindestens eines durch das Klassifikationsmodell hinreichend genau approximiert werden kann. Das Klassifikationsmodell ist wie im ersten Fall geeignet für die Beschreibung des Teilmusters. Nach der visuellen Separation muss der Nutzer entscheiden, ob die anderen Teilmuster durch das gleiche Modell beschrieben werden können, oder ob eventuell eine Kombination mehrerer unabhängiger Klassifikationsmodelle sinnvoll ist. Um dies zu beurteilen genügt es, das gleiche Klassifikationsmodell einzeln auf die separierten Teilmuster anzuwenden. Unter Umständen ist es sinnvoller, ein Muster als Kombination mehrerer einfacher Klassifikationsmodelle zu beschreiben als ein komplexes Modell zu verwenden.

Gerade am letzten Fall zeigt sich, dass der Anwender mit dem automatischen Verfahren in einen Dialog tritt, in dem Sinne nämlich, dass beide Seiten voneinander beeinflusst werden. In diesem Dialog kann durch geeignetes Hervorheben des modellierten Musters die Wahrnehmung des Bildes verändert werden, so dass es entweder zu einer Revision des ursprünglich wahrgenommenen Musters führt, oder zu einem Verwerfen des Modells bzw. des gesamten Verfahrens. Wenn sich Muster und Feedback wechselseitig positiv verstärken, ist der Dialog in dem Sinne erfolgreich, dass ein geeignetes Modell für das Muster gefunden wurde.
Das Kriterium, nach dem das wahrgenommene Muster und das Feedback verglichen werden, ist die Form der Punktmenge, d.h. insbesondere die Form ihres Randes. Wendet man verschiedene automatische Verfahren auf das gleiche interaktiv definierte Muster an, dann ist die Form dieser Punkte charakteristisch für Verfahren und Modell. Das Feedback liefert dem Nutzer gerade in den Fällen die meisten Informationen über das Prädiktionsmodell, in denen der Nutzer die Form des Rands durch die Interaktion nur indirekt beschreiben kann. In machen Ansätzen (siehe z.B. Liu et al. [LSG04]) wird der Rand des Prädiktors explizit beschrieben. In diesen Fällen unterscheidet sich das visuelle Feedback des Prädiktors prinzipiell nicht vom visuellen Feedback nach der direkten Selektion. Unter diesen Bedingungen wird das Prädiktionsmodell durch die Interaktion direkt und explizit bestimmt; es gibt keine Freiheitsgrade des Modells, die automatisch beschrieben werden könnten. Die manuelle Beschreibung ist in erster Linie deshalb problematisch, weil unklar bleibt, gemäß welcher Kriterien das Prädiktionsmodell das präziseste und/oder einfachste Modell innerhalb dieser Modellfamilie darstellen soll.
Für das Feedback kommen daher nur solche Verfahren in Frage

- in denen die Modellparameter *indirekt* - z.B. aus den vom Nutzer gegebenen Daten - berechnet werden oder

- in denen die Modellparameter die Punktemenge nur *implizit* beschreiben oder

- beides.

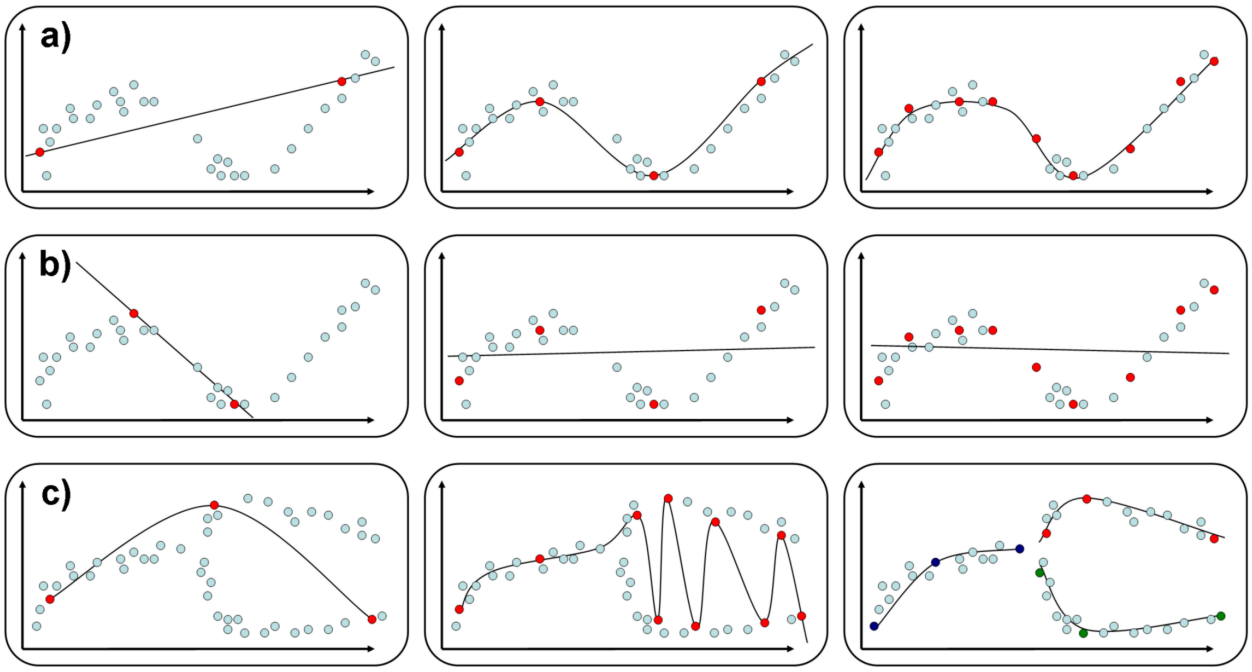


Abbildung 3.21: *Dieses Bild zeigt drei der vier Fälle, die als Ergebnisse des Abgleichs zwischen Mustererkennung und Feedback auftreten können. Exemplarisch wird hier eine Punktewolke in einem Scatterplot dargestellt, die durch ein Regressionsmodell approximiert werden sollen. Rot sind jeweils die durch den Nutzer selektierten Punkte. Das Modell selbst ist eine Kurve, die die selektierten Punkte approximiert. Fall (a) zeigt den Fall, dass wahrgenommenes Muster und Modell konvergieren. In diesem Fall ist das Verfahren also geeignet, um die Struktur des Datensatzs zu beschreiben. In Fall (b) ist das Modell - eine Regressionsgerade - zu einfach, um die Struktur des Datensatzes zu beschreiben. Ein Indikator dafür ist die Instabilität des Prädiktors (linkes und mittleres Bild) bzw. die Konvergenz gegen eine Lösung, die die Struktur des Datensatzes nicht erfasst (rechtes Bild). Fall (c) zeigt eine Struktur, für die ein einziges Regressionsmodell eine zu starke Vereinfachung darstellt (mittleres Bild) die in einem ungeeigneten Kompromiss endet. Durch die mehrfache Anwendung des Verfahrens auf Teilsegmente wird dieses Problem gelöst. (rechtes Bild).*

Zur ersten Gruppe gehören beispielsweise Verfahren für die *Intervallklassifikation*, in der die Form der Punktmenge explizit als n-dimensionales Intervall beschrieben ist und lineare Klassifikatoren und Regressionsmodelle, in denen die Punktmenge eines Musters durch einen linearen Unterraum beschrieben wird. Ebenfalls zu dieser Gruppe gehören die *Entscheidungs- und Regressionsbäume*, die auf den vorgenannten Verfahren aufbauen.
Ein einfaches Verfahren aus der zweiten Gruppe ist die *(k)-Nearest-Neighbor-Klassifikation*, in dem die Form der Punktmenge durch die interaktive definierten Trainingsdaten in Kombination mit der Abstandsfunktion auf dem Merkmalsraum beschrieben wird. Die Modellparameter sind die Samplepunkte, die direkt aus den Eingaben abgeleitet werden. Die Abstandsfunktion kann ebenfalls als Modellparameter angesehen werden, diese kann außer in trivialen Fällen jedoch nicht explizit durch den Nutzer beschrieben werden (siehe Abschnitt 3.1.4.6).
Zur dritten Gruppe gehören Verfahren wie beispielsweise der *k-Means-Algorithmus*. Der k-Means ist eigentlich ein Clusteringverfahren, jedoch kann die Suche nach Clustern durch die direkte Eingabe potentieller Clusterzentren erheblich verbessert werden. Im Unterschied zum einfachen *Nearest-Neighbor-Klassifikator* können die Clusterzentren beim k-Means zunächst

iterativ an die Trainingsdaten angepasst, und dabei ggf. verändert werden. Danach kann das Verfahren verwendet werden wie ein *Nearest-Neighbor-Verfahren.*
In allen relevanten Prädiktionsverfahren gibt es daher mindestens eine Indirektion bei der Verarbeitung der Eingaben des Nutzers. Diese Indirektion charakterisiert die Heuristik des Verfahrens, Gütefunktionen oder andere (implizite) Modellparameter. Daher folgt daraus, dass das visuelle Feedback im allgemeinen *nicht* automatisch deckungsgleich ist mit dem wahrgenommenen Muster. Die Abweichungen zwischen Muster und Feedback und deren Entwicklung während der Arbeit mit dem Muster sind das Kriterium für die Qualität des Prädiktors bzw. des Verfahrens.
Zwei Dinge sind zu betonen: Erstens handelt es sich bei durch die automatischen Verfahren konstruierten Modelle um mathematische Beschreibungen, die unabhängig sind von der verwendeten Visualisierungstechnik. Die sichtbare Form der Punktmenge hängt natürlich von der Visualisierung selbst ab. Insbesondere muss das durch die Klassifikation bestimmte Muster nicht zwangsläufig durch einen räumlich zusammenhängenden Bereich dargestellt werden. Wenn beispielsweise das für das Muster wichtigste unabhängige Attribut des Merkmalsraums nicht auf die Position, sondern auf die Farbe abgebildet wird, wird beim visuellen Feedback eher ein Bereich markiert, der sich durch eine ähnliche Farbgebung hervorhebt. Dieses Feedback ist sogar erwünscht, wenn das ursprünglich durch die Interaktion definierte visuelle Muster über die Farbe charakterisiert wurde. Prinzipiell ist das visuelle Feedback auf alle verwendeten visuellen Attribute einer Visualisierung anwendbar.
Zweitens folgt daraus, dass die Kenntnis über die Mathematik der Prädiktionsverfahren nur bedingt bei der Bewertung dieser Verfahren hilft. Man müsste nicht nur in der Lage sein, die Punktmenge zu beschreiben, die ein Prädiktormodell im Muster zusammenfasst (sofern das überhaupt möglich ist), sondern zusätzlich noch deren Abbild in der jeweils verwendeten Visualisierung. Bei einem entsprechenden Repertoire an wählbaren Methoden wären dabei auch Experten überfordert. Dies mag als Nachteil erscheinen, die Kernidee besteht stattdessen aber gerade darin, die Maschine dieses gekoppelte Korrespondenzproblem lösen zu lassen (siehe Abbildung 3.22), das entsteht, wenn Visualisierungs- und Data-Mining-Technik miteinander verbunden werden. Wie eingangs beschrieben, ist die Qualität eines Prädiktors davon abhängig, wie genau und wie effizient er Beziehungen innerhalb des Merkmalsraums reproduziert. Anhand dieser Eigenschaften kann er durch Interaktion und Feedback überprüft werden, wobei für den visuellen Abgleich nicht bekannt sein muss, wie die Verfahren die Korrespondenz herstellen.
Das visuelle Feedback auf der Basis von Prädiktormodellen kann als Verallgemeinerung der *Linking & Brushing* Techniken aufgefaßt werden, die für die Modellierung von Suchanfragen konzipiert werden (siehe auch Abschnitt 2.4.3.2). Die Abarbeitung der Queries in der Datenbank entspricht konzeptionell der Anwendung des Prädiktors auf die gegebenen Daten. Da die Ergebnisse des Prädiktors in jeder Visualisierung dargestellt werden können, die eine oder mehrere Attribute des Merkmalsraums darstellt, können auf diese Weise verschiedene Visualisierungen miteinander gekoppelt werden.

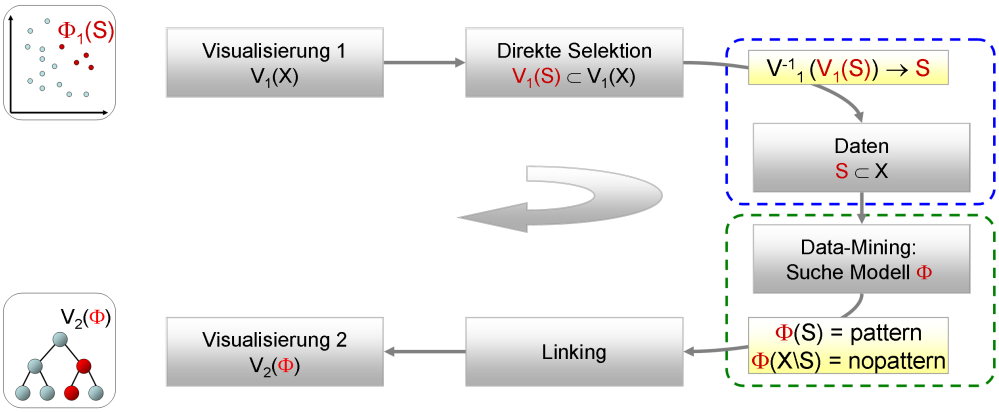


Abbildung 3.22: *Die Verbindung zwischen Datenvisualisierung und Modellvisualisierung kann auch als Verallgemeinerung des* Linking & Brushing *aufgefasst werden (siehe auch Abbildung 2.15). Beim Brushing ist das Datenmodell die für beide Visualisierungen V_1 und V_2 gemeinsame Referenzrepräsentierung (blau). Eine Selektion wird beim Brushing daher in einem Datenmodell beschrieben. Die Musterbeschreibung ist in diesem Sinne lediglich eine weitere Transformation dieses Datenmodells in ein analytisches Modell (grün). Brushing läuft in beide Richtungen gleich ab. Mit dieser Erweiterung wird der umgekehrte Prozess dadurch definiert, dass das Data-Mining Verfahren durch eine Simulation ergänzt wird.*

3.2.3.2 „Fuzzy“ Feedback und Vermeidung von Überanpassung

Beim visuellen Feedback wurden bisher nur die Fälle berücksichtigt, in denen der Prädiktor eine eindeutige Klassenzuordnung liefert. Dementsprechend lieferte das visuelle Feedback eine binäre Unterteilung des Darstellungsraums in ein Muster bzw. ein Nicht-Muster. In zwei Fällen erscheint es sinnvoll, eine feinere Abstufung zuzulassen:
Der erste Fall sind probabilistische Verfahren, die keine scharf begrenzten Muster als Feedback für die Klassifikation liefern, sondern eine Dichtefunktion für jede Klasse. Bei der Bestimmung des maximalen Erwartungswertes, der die Kategorie bestimmt, geht Information verloren. Durch das unscharfe Feedback kann man die Dichtefunktion direkt darstellen. Zusätzlich werden alle Bereiche der Dichtefunktion, die durch das interaktiv definierte Muster beeinflusst werden, dargestellt. Scheinbar unwichtigere Bereiche sind vor dem Hintergrund relevant, weil das Feedback dem Anwender auch Anhaltspunkte darüber geben soll, wo er durch weitere Samples den Klassifikator besonders effektiv verfeinern kann. Das verwendete Ergebnis des probalistischen Verfahrens ist nicht der Prädiktor, sondern die Wahrscheinlichkeit; dementsprechend muss für das Feedback ein visuelles Attribut verwendet werden, das in jedem Fall für numerische Werte geeignet ist.
Der zweite Fall betrifft Prädiktionsverfahren, für die ein Parameter frei wählbar ist, der den „Trade-Off“ zwischen Approximationsgenauigkeit der Trainingsdaten und der Komplexität des Modells bestimmt (siehe *Occams Dilemma* in Abschnitt 2.3.3). Dieser Trade-Off ist unabhängig von Trainingsdaten und Abstands- oder Gütefunktionen innerhalb des Merkmalsraums. Zum wichtigsten Vertreter dieser Klasse zählen die Entscheidungs- bzw. Regressionsbäume. In diesem Fall soll das visuelle Feedback keine binären, sondern abgestufte Informationen darüber liefern, zu welchem Maße ein Punkt bzw. ein Objekt des Merkmalsraums bereits zum modellierten Muster gehört oder nicht. Konkret kann so jedem Punkt ein Maß zugeordnet werden, das angibt, wie sehr das Hinzufügen eines Punktes zum selektierten

Muster die Komplexität des Modells erhöht.
Im Gegensatz zu den Modellen, in denen die Modellkomplexität direkt oder indirekt durch die Anzahl der Samples beschrieben werden kann (z.B. *Nearest-Neighbor-Klassifikation*) oder grundsätzlich beschränkt ist (z.B. *Regressionsgeraden oder -Polynome vorgebenen Grades*), kann ein geeigneter Wert für diesen Trade-Off nicht durch Interaktion mit einem Muster bestimmt werden. Die Prädiktionsverfahren dieser Klasse können die gegebenen Trainingsdaten daher mit beliebigen Aufwand und beliebiger Genauigkeit approximieren. Jeder Wert für diesen Trade-Off kann die Trainingsdaten auf unterschiedliche Weise zu Punktmengen zusammenfassen. Daher besteht die Bedeutung eines Parameters für diesen Trade-Off im Kern darin, abzugrenzen, welche Samples zum Muster zusammengefasst werden und welche Samples als Rauschen ignoriert werden sollen. Dass diese Grenze flexibel und ohne grundsätzliche Veränderung des Prädiktormodells verschoben werden kann, ist die eigentliche Stärke dieser Verfahren. Wie bei allen automatischen Verfahren gilt jedoch auch hier, dass es keine Garantie dafür gibt, dass ein beliebig gewählter Parameterwert gerade die Punktmenge zu dem Muster zusammenfasst, die ein Anwender in einer Visualisierung vom Rauschen separiert.
Um die Überanpassung zu vermeiden, sind verschiedene automatische und interaktive Strategien denkbar:

- Automatische Verfahren wie *Bootstrap* Methoden und/oder mehrfache *Kreuzvalidierung* [Bre01a, Koh95] umgehen diesen Trade-Off dadurch, dass mehrere Prädiktormodelle auf unterschiedlichen Trainingsdaten erzeugt werden, wobei das Endergebnis gemittelt, bzw. abgestimmt wird. Allerdings multipliziert sich dadurch die Modellkomplexität mit der Zahl der Modelle. Ob diese Komplexität im nachhinein vereinfacht werden könnte, würde selbst eine eigene Analyse auf den Modellparametern erfordern.

- Die halbautomatische Bestimmung des Verfahrensparameters bei der Konstruktion des Modells. Dazu gehören beispielsweise Verfahren, mit denen bei Entscheidungs- und Regressionsbäumen eine Überanpassung vermieden werden soll (sogenannte *Pruningstrategien*). Ihnen gemeinsam ist, dass die Genauigkeit der Approximation in Abhängigkeit von der Modellkomplexität bestimmt wird. In den Strategien wird die Komplexität angepasst an verschiedene Maßzahlen, wie zum Beispiel die gewünschte Genauigkeit oder der gewünschte Grenznutzen für die Erhöhung der Komplexität um die nächste Stufe. Schaffer [Sch93] legt jedoch dar, dass *jede* Form einer automatischen Strategie Annahmen über das Modell birgt, das die Daten geeignet beschreiben kann. Solange diese Annahmen nicht bestätigt werden, stünde diese Strategie im Widerspruch zu diesem Konzept. Dies gilt gleichermassen für die Wahl eines Kriteriums wie auch für die Wahl der Komplexität.

- Die Variante der halbautomatischen Verfahren, die hier vorgestellt wird, ist die *interaktive* Bestimmung dieses Parameters durch visuellen Abgleich. Unter der Voraussetzung, dass das Prädiktionsmodell hinreichend schnell berechnet werden kann, kann der Anwender diesen Parameter steuern und die Feedbackfunktion in der Visualisierung nutzen, um ihn so lange einzustellen bis das Feedback mit dem Muster übereinstimmt. In diesem Fall wird der Parameter für die Komplexität des Modells durch einen visuellen Abgleich mit den Daten bestätigt.

Die Strategie für die interaktive Steuerung der Komplexität des Modells erfordert sowohl die direkte Interaktion innerhalb der Visualisierung für die Definition des Musters, als auch eine indirekte Interaktion für die Steuerung der Komplexität des Modells. Allerdings wird dabei zu jedem Zeitpunkt jeweils das Feedback des Prädiktormodells für einen bestimmten Komplexitätswert dargestellt. Stattdessen kann die Variation des Feedbacks über die verschiedenen Komplexitätswerte dargestellt werden. Die obere Grenze für die Komplexität wäre das Modell, das die Klassifikation der durch den Nutzer gegebenen Samples fehlerfrei beschreibt. Technisch kann diese Variation als Überblendung mehrerer visueller Feedbacks gelöst werden. Durch die Überblendung verschiedener Muster erhält der Anwender Informationen über

- die maximale Menge von Punkten, die mit den aktuellen Verfahren unabhängig von der Modellkomplexität zum Muster gehören können,
- die globale und lokale Stabilität des Prädiktormodells in Abhängigkeit von seiner Komplexität

Die Stabilität eines Modells bezieht sich darauf, welche Ähnlichkeit das visuelle Feedback von Modellen unterschiedlicher Komplexität miteinander hat. In beiden Fällen handelt es sich um einen subjektiven Eindruck, der beschreibt, ob der Anwender Regelmäßigkeiten zwischen den Bildern der Modelle unterschiedlicher Komplexität erfassen kann. Globale Stabilität beschreibt die Eigenschaften, die über alle Komplexitätsstufen und alle erzeugen Modelle gleich sind. Diese Eigenschaften charakterisieren die Modellfamilie des Verfahrens. Ein Beispiel für solche Eigenschaften ist die Tatsache, dass die Grenzen zwischen Klassen in einem univariaten Entscheidungsbaum stets entlang der Hauptachsen des Merkmalsraums verlaufen oder dass sich die Grenzen zwischen Klassen beim k-Nearest-Neighbor Verfahren stets aus Teilen von Hyperebenen zusammensetzen (siehe Abbildung 2.7 und 3.23).
Sofern eine Visualisierung diese Eigenschaften überhaupt darstellt, kann der Anwender überprüfen, ob das wahrgenommene Muster diese ebenfalls erfüllt. Sind diese Eigenschaften nicht übertragbar auf das wahrgenommene Muster, dann ist es wahrscheinlich, dass Modell und Verfahren für dieses Muster verworfen werden müssen.
Lokal stabile Bereiche des Musters sind solche, in denen der Prädiktor über einen großen Komplexitätsbereich die gleichen Ergebnisse liefert. Diese werden durch das Überzeichnen hervorgehoben. Wenn der Anwender ein neues Sample definiert, um die Eingabedaten für das Prädiktorverfahren zu erweitern, verstärkt er implizit genau die Komplexitätsstufen des Prädiktormodells, die die Samples richtig beschreiben. Durch die Überzeichnung kann abgeschätzt werden, welche Auswirkungen ein neues Sample haben wird. Verstärkt beispielsweise der Nutzer durch ein Sample einen Bereich, der lokal stabil ist, und daher auf den meisten Komplexitätsstufen sowieso bereits den Wert des Samples reproduziert, dann kann der Nutzer erwarten, dass sich Modell und Feedback nur geringfügig ändern. Unterscheidet sich der Wert des Samples in einem lokal stabilen Bereich vom Prädiktionswert des Modells, kann der Anwender eine Änderung von Modell und Feedback erwarten. Durch die Definition eines Samples in einem lokal instabilen Bereich wird das Modell am wirkungsvollsten verfeinert, und der Komplexitätsbereich des Prädiktors, in dem das wahrgenommene Muster modelliert wird, wird am besten eingeschränkt.

Die Überzeichnung bietet daher gegenüber der interaktiven Steuerung der Komplexität im GUI und des Feedbacks nur eines Musters (bzw. eines Komplexitätsgrades) zwei Vorteile:

- Der Anwender kann gleichzeitig mehrere Komplexitätsstufen des Prädiktors miteinander vergleichen.
- Der Anwender erhält genauere Informationen darüber, wie das Modell auf weitere Samples reagiert und kann daher auch die Stabilität eines Prädiktors abhängig von gegebenen Eingabedaten besser einschätzen.

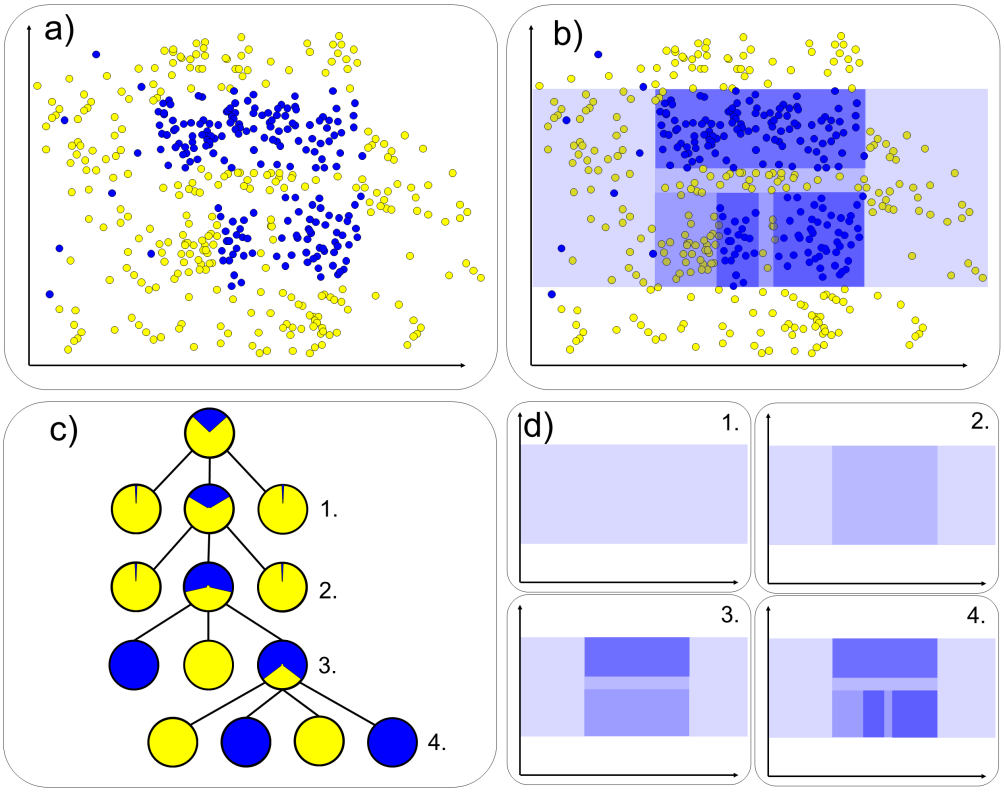


Abbildung 3.23: *Durch die gleichzeitige Darstellung verschiedener Komplexitätsstufen von Prädiktoren kann der Anwender die Stufe identifizieren, die dem wahrgenommenen Muster am nächsten kommt. Bild (b) zeigt das Feedback eines Prädiktors, der durch einen Entscheidungsbaum definiert wurde. Das Bild der Muster im (univariaten) Entscheidungsbaum ist stets eine Menge achsenparalleler Rechtecke im Merkmalsraum. Jedem Punkt im Darstellungsraum wird ein Wert zugeordnet, der abhängig davon ist, welchen Wert der Klassifikator in seinen verschieden komplexen Varianten für diesen Punkt annimmt (c) + (d). Je nachdem, welche Bereiche der Anwender in der Folge selektiert oder deselektiert, kann eine bestimmte Komplexitätsstufe für den Klassifikator gewählt werden.*

Die Überzeichnung ist formal der gewichtete Durchschnitt aller Werte, die der Klassifikator auf verschiedenen Komplexitätsstufen für den gleichen Punkt des Merkmalsraums annimmt. Zwischen der minimalen und der maximalen Komplexität des Modells wird der Anteil der Klassifikatoren bestimmt, die einen Punkt x den Wert $\Phi(x) = pattern$ enthalten. Wie dieser Anteil berechnet wird, hängt von der Art der Klassifikatoren ab. Bei Entscheidungsbäumen kann dieser Anteil beispielsweise über den *Information-Gain* berechnet werden (siehe

Quinlan [Qui96]). Dabei handelt es sich um ein Entropiemaß, das jedem Knoten zugeordnet werden kann, und das bestimmt, welches Attribut in diesem Knoten verwendet wird. Der Wert, den der Entscheidungsbaum beim Pruning an diesem Knoten haben würde, wird mit dem Entropiemaß gewichtet. Der gewichtete Durchschnitt an einem Punkt des Merkmalsraums wird über alle Knoten vom Blatt- bis zum Wurzelnknoten bestimmt. Im folgenden Kapitel wird für das Feedback ein Verfahren vorgestellt, das diese Strategie für die Bewertung von Entscheidungsbäumen umsetzt (siehe Abschnitt 4.1.4).
Bei der gewünschten Komplexität des Klassifikators handelt es sich um einen Verfahrensparameter. Durch das „Fuzzy"-Feedback kann überprüft werden, ob die Lösung stabil bezüglich dieses Parameters ist, und durch die Wechselwirkung zwischen Feedback und Interaktion kann der Anwender innerhalb der Visualisierung eine bestimmte Komplexitätsstufe wählen. Der Thresholdparameter, der bei Entscheidungsbäumen das Pruning steuert, kann so auf der Grundlage eines Musterabgleichs gewählt werden. Dass es einen solchen Verfahrensparameter gibt, ist jedoch die Voraussetzung für diese Strategie.
Nicht untersucht wurde im Rahmen dieser Arbeit, ob diese Strategie auch übertragen werden kann, um auch andere Verfahrensparameter durch visuellen Abgleich zu bestimmen. Beispielsweise wäre es denkbar, dass die Anzahl der Nachbarn k bei der *k-Nearest-Neighbor* Klassifikation in ähnlicher Weise untersucht und auch interaktiv definiert werden könnte. Eine Herausforderung bleibt in jedem Fall die gleichzeitige Untersuchung mehrerer Parameter, da das Feedback die Variationen der Modelle nur entlang eines Parameters beschreibt.

3.3 Synthese und Zusammenfassung

In diesem Kapitel wurden mehrere Kopplungsvarianten für die Verbindung zwischen visuell-interaktiven und automatischen Verfahren für die Datenanalyse vorgestellt. Im ersten Teil des Konzepts wurde die Interaktion des Menschen in einer Visualisierungstechnik als Datenquelle für die Steuerung automatischer Verfahren genutzt. Dieses Konzept steht komplementär zu den Kopplungsmodellen, in denen die Visualisierung gewissermaßen die Datensenke darstellt. Dazu zählen Verfahren, in denen die Visualisierung dazu dient, die Ergebnisse von automatischen Verfahren darzustellen, aber auch Verfahren, in denen die Visualisierungstechniken automatisch gesteuert werden. In diesem Sinne soll dieses Konzept das Repertoire möglicher Kopplungen weiter vervollständigen.

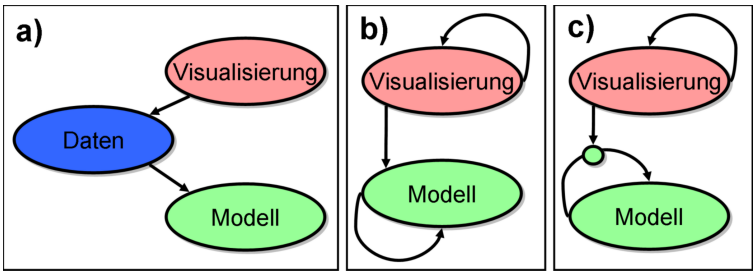


Abbildung 3.24: *Basierend auf dem Modell von Fayyad et al. (siehe Abschnitt 2.3.4) wurden drei Ansatzpunkte für eine Steuerung der automatischen Verfahren durch Visualisierungstechniken identifiziert. Die erste und in diesem Konzept wichtigste Art der Kopplung ist die manuelle Definition von Eingabedaten für die Analyse über die direkte Selektion in der Visualisierung (a). Die zweite Variante ist die manuelle, direkte Definition von Modellparametern (b). Diese Variante wurde in existierenden Techniken bereits am häufigsten umgesetzt. Die letzte Variante beschreibt die Steuerung des automatischen Verfahrens über die Darstellung der Verfahrensparameter (c). Im Gegensatz zu der häufig eingesetzten Steuerung über GUI-Element wird die Steuerung aus der Visualisierung am seltensten umgesetzt.*

Auf der Grundlage des allgemeinen Modells von Fayyad et al., das im vorigen Kapitel vorgestellt wurde (siehe Abschnitt 2.3.4), wurden drei Ansatzpunkte für eine Nutzung dieser Daten in einem automatischen Verfahren identifiziert. Der erste und im Rahmen dieses Konzepts wichtigste der drei Ansatzpunkte ist die Interpretation der Daten, die aus der Interaktion stammen, als Eingabedaten für das automatische Verfahren. Auf dieser Art von Kopplung gründet sich die Separation von Mustererkennung (durch den Menschen) und Musterbeschreibung (die Modellierung durch das automatische Verfahren). Diese Trennung ist motiviert durch die unterschiedlichen Stärken von Mensch und Maschine.
In der Informationsvisualisierung muss der Mensch diese Aufgaben gleichermaßen erfüllen. Die Stärke des Menschen liegt aber viel eher im Bereich der Erkennung von Mustern und Strukturen in Bildern. Die Beschreibung eines wahrgenommenen Musters ist notwendig, um seinen Informationsgehalt im Kontext der Aufgabenstellung der Analyse interpretieren und evaluieren zu können. Sie erfordert jedoch die Aufmerksamkeit des Nutzers und damit kognitiven Aufwand. Dabei entsteht das Problem, dass die Komplexität der beschreibbaren Inhalte durch die kognitiven Kapazitäten des Menschen beschränkt ist.
Umgekehrt besteht in der automatischen Analyse das kritischste Problem darin, dass mit der

Wahl eines Verfahrens die Menge der Muster determiniert ist, die mit hinreichend einfachen - und damit aussagekräftigen - Modellen noch zu beschreiben ist. Man kann insbesondere in der explorativen Analyse nicht davon ausgehen, dass die Wahl eines Verfahrens immer richtig ist im Bezug auf Daten und Aufgabe. Auch ist die Komplexität der Heuristiken automatischer Verfahren ein Indiz dafür, dass automatische Suche nach Mustern ein entsprechend schwieriges Problem darstellt.
Die Aufteilung von Mustererkennung und Musterbeschreibung wird daher den spezifischen Stärken von Mensch und Maschine gerecht. Die vorgeschlagene Kopplung, die direkte Selektion von Mustern und deren Interpretation innerhalb eines automatischen Verfahrens, wurde dabei besonders unter dem Gesichtspunkt entworfen, den kognitiven Aufwand für den Menschen möglichst gering zu halten. Auf der Grundlage der Interaktionsmodelle von Lam und Norman (siehe Abschnitt 2.4.3) wurde dieser kognitive Aufwand als Korrespondenzproblem zwischen Wahrnehmung und Handlung formuliert. Nur eine direkte Selektion von Mustern oder allgemeiner Bildbereichen stellt eine Form der Interaktion dar, in der es eine direkte räumliche Korrespondenz zwischen Muster und Reaktion gibt.
Abhängig von der Visualisierungstechnik ist es möglich, den selektierten Bereich als Teilmenge des dargestellten Datensatzes oder des dargestellten Merkmalsraums zu interpretieren. Durch die Selektion zeichnet der Mensch diese Teilmenge als Muster aus und liefert dem automatischen Verfahren damit Informationen, für die sonst eine automatische Suche notwendig wäre. Die automatischen Verfahren setzen die Verarbeitung mit der Beschreibung bzw. Modellierung dieses Musters fort. Die Trennung von wahrgenommen Muster und Nicht-Muster ist formal ein Klassifikationsproblem. Es wurde jedoch auch gezeigt, dass die Kombination von Mustererkennung und -beschreibung ebenso auch auf das Clusteringproblem angewendet werden kann.
Die automatische Musterbeschreibung hat abgesehen von der Entlastung des Menschen noch weitere Vorteile gegenüber der Interpretation durch den Menschen: Zum einen ist sie ein deterministischer, nachvollziehbarer Prozess, dessen Ergebnisse formal beschrieben sind, und die maschinell weiterverarbeitet werden können. Erst die Reproduzierbarkeit von Ergebnissen macht diese überhaupt wertvoll; durch die automatische Interpretation werden die Ergebnisse der Wahrnehmung dem analytischen Diskurs exponiert.
Zum anderen sind automatische Verfahren nicht auf die dargestellten Daten beschränkt. Bei der Modellierung können wesentlich mehr Daten und Attribute berücksichtigt werden, als eine Visualisierung dem Menschen zugänglich machen kann. Durch das größere Repertoire an Daten können zudem Verknüpfungen zwischen dem selektierten Muster und nicht sichtbaren Attributen identifiziert werden, die in den folgenden Iterationen auch wieder visuell überprüft werden können.

Der zweite Ansatzpunkt für die Interpretation der Daten ist die Interpretation der Eingabedaten einer Visualisierung als Modellparameter der automatischen Verfahren. Diese Variante der Kopplung wurde bisher am häufigsten in existierenden Ansätzen umgesetzt. In diesen Ansätzen können die Modellparameter visuell editiert werden. Auch diese Ansätze basieren auf der direkten Selektion innerhalb der Visualisierung.
Bei der Interpretation als Modellparameter können Verfahren danach unterschieden werden, ob und wie viele Parameter durch den Anwender direkt beschrieben werden, oder ob sie für die Heuristik initialisiert werden. Wesentlich unterscheiden sich die Verfahren vom hier

vorgestellten Konzept, wenn *alle* Modellparameter durch den Anwender *direkt* beschrieben werden. Eine Kopplung mit einem automatischen Verfahren findet dabei nicht statt.
Im Abschnitt 3.1.6 wurden die Funktion dieser Kopplungsvarianten auch aus der Perspektive automatischer Verfahren allgemein beschrieben. Die Interpretation der Interaktion als Eingabedaten verändert dabei indirekt die Eigenschaften der Gütefunktion. Die Gütefunktion wird insofern vereinfacht, da sich die Optimierung auf die Struktur und den Detaillierungsgrad fokussiert, die der Anwender in der Interaktion angegeben hat. Dass die Heuristik in einer Kompromisslösung konvergiert, wenn mehrere Strukturen nicht gleichzeitig „gut“ beschrieben werden können, wird auf diese Weise vermieden.
Die Interpretation der Interaktion als Modellparameter verkleinert den Suchraum für die Heuristik entweder explizit, wenn die Parameter direkt gesetzt werden, oder implizit, wenn die Parameter initialisiert werden. Dies gilt jeweils unter der Bedingung, dass die optimale Lösung durch den Anwender in der Visualisierung approximiert werden kann.
Die Kopplung über die Eingabedaten ist unabhängig von Visualisierungstechnik und automatischem Verfahren. Die Kopplung über die Modellparameter ist abhängig vom analytischen Modell. Der dritte Ansatzpunkt für die Kopplung ist die Steuerung des automatischen Verfahrens selbst über die Manipulation seiner Verfahrensparameter. Verfahrensparameter werden nur selten über eine Visualisierungstechnik definiert, sondern meist über andere Komponenten der Nutzerschnittstelle geändert.
Dennoch erscheint es sinnvoll, die Auswahl von Verfahrensparametern als Interaktion innerhalb einer Visualisierung zu gestalten. Dahinter steht die Annahme, dass die Wahl des Wertes für einen beliebigen Verfahrensparameter oft ein multikriterielles Entscheidungsproblem darstellt, und dass die Wahl das Ergebnis dieses Teilprozesses wesentlich beeinflusst. Eine methodische Analyse erfordert überdies, dass die Beziehung zwischen dieser Wahl und dem Ergebnis der Analyse so gut wie möglich exponiert wird. Eine Visualisierung erlaubt es - im Unterschied zu anderen Steuerelementen in der Graphischen Nutzerschnittstelle - die Entscheidung im Kontext der dafür relevanten Kriterien zu treffen. Mögliche Kriterien wurden in Abschnitt 3.1.4.3 genannt.
Im Abschnitt 2.3.2 wurde dargelegt, das die automatische Analyse in hohem Maße von vorbereitenden Prozessen abhängt. Eine dieser Vorbereitungen ist die geeignete Selektion der Attribute eines Merkmalsraums, die für die weiteren Schritte der Analyse genutzt werden sollen. Für diese Aufgabe wurde exemplarisch dargelegt, wie eine Kopplung zwischen Visualisierung und Verfahrensparametern aussehen kann. Im Unterschied zu den anderen beiden Kopplungsvarianten ist es dafür notwendig, die Visualisierung an dieses Entscheidungsproblem anzupassen.
Bei der Umsetzung dieser Kopplungsvariante wurde vorausgesetzt, dass die für die Wahl von Verfahrensparametern relevanten Kriterien von vornherein bekannt sind, so dass in der Visualisierung nur die Informationen konsolidiert werden müssen, die im konkreten Entscheidungsfall vorliegen. Denkbar, jedoch im Rahmen dieses Konzepts nicht behandelt, ist stattdessen auch der Fall, dass auch die Kriterien für diese Wahl nicht bekannt sind. In diesem Fall rechtfertigt die Wahl der Verfahrensparameter eine explorative Analyse eigenen Rechts. Hierzu sei auf den Ausblick der Arbeit verwiesen.

Der zweite Teil dieses Konzepts beschreibt die Zusammenführung von konfirmativer und explorativer Datenanalyse, wie sie im Zusammenhang mit der Kopplung zwischen visuell-interaktiven und automatischen Verfahren umgesetzbar ist. Für diese Zusammenführung

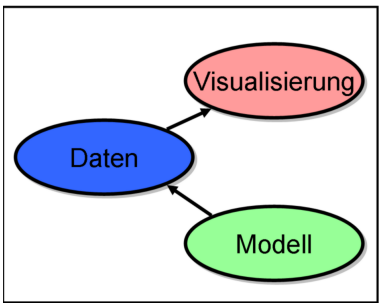


Abbildung 3.25: *Der zweite Teil des Konzepts beschreibt den komplementären Prozess, in den konfirmative und explorative Analyse verbunden werden. Dabei wird ein prädiktives Modell durch Simulation auf Rohdaten abgebildet, die wiederum als Datenquelle für einen visuellen Abgleich zwischen simulierten Daten und Referenzdaten genutzt werden. Diese Kopplung dient als Feedback für das Verfahren, mit dem das Modell konstruiert wird.*

wurden zwei verschiedene Varianten identifiziert.
In Abschnitt 3.2.2 wurde allgemein die explorative Analyse beschrieben als Möglichkeit, die Fehler der Modell- und Hypothesenbildung nicht nur quantitativ zu beschreiben, sondern auch qualitativ. Ein systematischer Fehler bei der Modellierung kann sich beim visuellen Abgleich als Muster manifestieren. Ein prädiktives Modell wird durch einen Simulationsprozess transformiert auf das Datenmodell der Rohdaten. Simulierte Daten und Rohdaten werden in der gleichen Visualisierung einander gegenübergestellt, so dass ein systematischer Fehler durch den Nutzer identifiziert werden kann.
Dieses Konzept ist für jedes prädiktive Modell anwendbar, unabhängig davon, ob es von einer Hypothese abgeleitet wurde, oder ob es durch eine Analyse erst erzeugt wurde. Im Abschnitt 3.2.3 wurde beschrieben, wie dieser visuelle Abgleich integriert wird in den Prozess von Mustererkennung und Musterbeschreibung, der im ersten Teil dieses Konzepts entwickelt wurde.
Das visuelle Feedback wird als Erweiterung des Feedback verstanden, mit dem eine Visualisierung nach einer direkte Selektion die selektierten Objekte oder Bereiche hervorhebt. Dadurch, dass dieses Feedback nicht direkt, sondern über ein automatisches Analyseverfahren bestimmt wird, kann der Modellierungsfehler aus dem Feedback abgelesen werden. Durch das iterative Feedback (siehe Abschnitt 3.2.3.1), bei dem das Modell nach jeder Interaktion neu berechnet wird, wechselwirken die Wahrnehmung der Muster durch den Menschen und die automatische Modellierung. Je nachdem, welche Muster das automatische Verfahren beschreiben kann, endet dieser Prozess entweder in der visuellen Konvergenz zwischen wahrgenommenem und beschriebenem Muster, oder damit, dass der Mensch die Anwendung des automatischen Verfahrens in der vorliegenden Konfiguration verwerfen kann.
In Abschnitt 3.2.3.2 wurde das Feedback verfeinert, um nicht nur eine binäre Klassifikation zwischen Muster und Nicht-Muster zu ermöglichen, sondern um bei der Unterteilung verschiedene Abstufungen zuzulassen. Diese Abstufungen können einerseits beim Feedback für probabilistische Modelle genutzt werden. Anderseits ist es zweckmäßig, sie auch bei solchen Modellen zuzulassen, in denen die Grenze zwischen Muster und Rauschen von der gewünschten Komplexität des Modells abhängt. Durch die Abstufung wird so nicht nur dargestellt, ob ein beliebiges Objekt im Merkmalsraum eher zum selektierten Muster gehört oder nicht.

Das System kann durch die Verfeinerung auch eine Rückmeldung darüber geben, ob die Hinzunahme eines Objektes zur Selektion die Komplexität des Modells erhöht oder nicht.

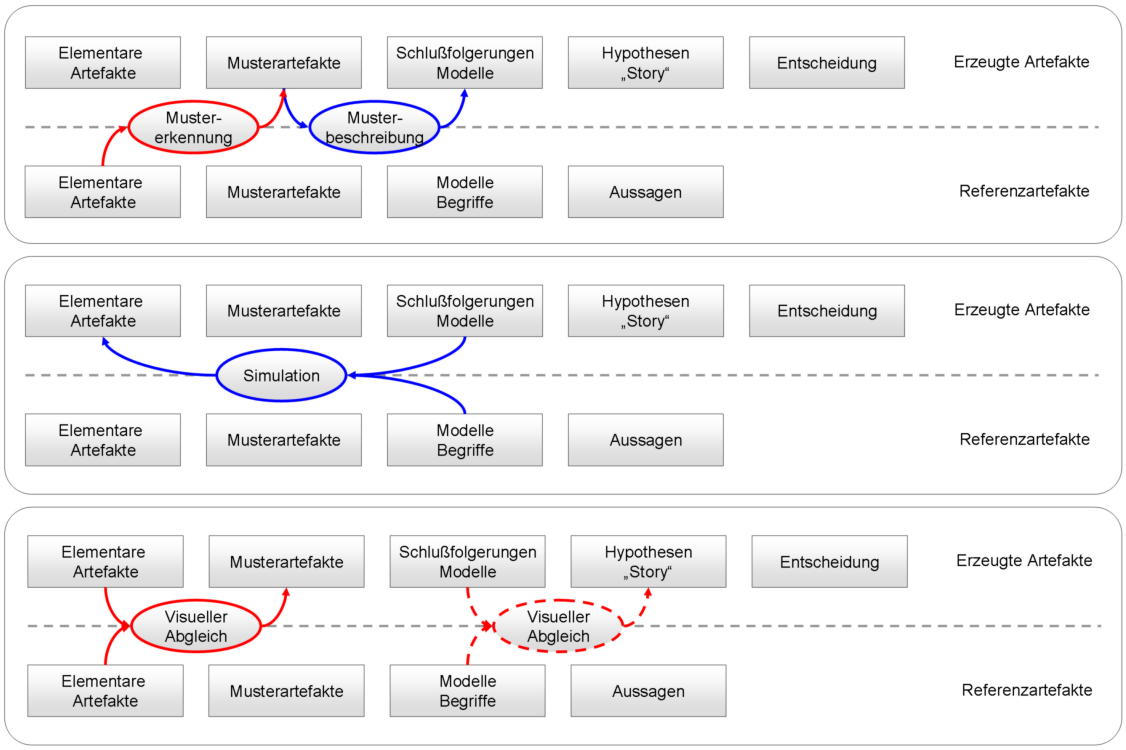


Abbildung 3.26: *Die drei Hauptziele dieser Arbeit werden mit Hilfe der Kopplung von automatischen Verfahren und visuell-interaktiven Verfahren umgesetzt. Die Transformation von elementaren Artefakten über Muster hin zu Modellen in zwei Schritten, die die Trennung zwischen Mustererkennung und Musterbeschreibung darstellen (oben), widerspiegelt das erste Ziel. Das zweite Ziel entspricht der Transformation von prädiktiven Modellen zurück auf elementare Artefakte über eine Simulation (mitte) mit dem Ziel, auf dieser Abstraktionsebene einen visuellen Abgleich zu ermöglichen (unten). Um das dritte Ziel zu erreichen, werden diese beiden Prozesse in einem Zyklus kombiniert. Die komplementären automatischen Verfahren können so über diesen Abgleich gegeneinander validiert werden.*

Das in dieser Arbeit vorgestellte Konzept beschreibt Methoden für die Transformation von elementaren Artefakten zu Mustern, von Mustern zu Modellen und zurück. Diesem Konzept liegen mehrere zentrale Gedanken zugrunde:
Wie in Abschnitt 2.1 beschrieben, müssen in einer Analyse Möglichkeiten dafür geschaffen werden, dass die Artefakte, die bei der Analyse konstruiert werden, gegen die Referenzartefakte mit dem jeweils gleichen Abstraktionsgrad getestet werden können. Abhängig davon, was die Referenzartefakte sind, erfordert dies die Transformation einzelner Artefakte in die Richtung eines höheren oder in die Richtung eines niedrigeren Abstraktionsgrades. Diese beiden Richtungen werden durch die Strategien der explorativen und der konfirmativen Analyse umgesetzt, die in diesem Konzept verbunden werden.
In dieser sehr starken Vereinfachung des Analyseprozesses lassen sich zwei fundamentale Operationen identifizieren. Die erste dieser Operationen ist der Vergleich zweier kompatibler Artefakte - also Artefakte, die die gleichen Phänomene auf dem gleichen Abstraktionsniveau

beschreiben. Die zweite Operation ist der Wechsel der Repräsentierung, wobei die Wechsel zwischen den verschiedenen Abstraktionsniveaus von besonderer Bedeutung sind.
In diesem Konzept wurde eine Arbeitsteilung zwischen Mensch und Maschine beschrieben, die sich auch über diese beiden Operationen beschreiben lässt. Aufgabe des Menschen ist erstens die Transformation von elementaren Artefakten in Musterartefakte. Zweitens gehört zu diesen Aufgaben der Abgleich auf der Ebene der Muster, wie auch auf der Ebene der Modelle. Wichtigstes Hilfsmittel für diese drei Aufgaben sind die Visualisierungstechniken. Die Visualisierung wird einerseits für die Erkennung von Mustern in den Daten genutzt, in besonderem Maße betont dieses Konzept aber auch die Möglichkeit, Muster in den Unterschieden zwischen konstruierten Artefakten und Referenzartefakten zu finden. Visualisierungstechniken ergänzen das Repertoire für den Abgleich von Artefakten und ermöglichen im Unterschied zu statistischen Fehlermaßen auch qualitative Aussagen über die Modellierungsfehler. Unter anderem erlauben sie Rückschlüsse darauf, ob ein Analyseverfahren für die Beschreibung eines Musters geeignet ist.
Die Aufgabe der automatischen Verfahren ist die Transformation von Mustern zu Modellen, bzw. von Modellen zurück zu Mustern oder elementaren Artefakten. Die Verfahren stellen für den Menschen potentiell die Korrespondenzen zwischen Mustern und Modellen her. Dies ist deshalb sinnvoll, weil sich diese Korrespondenz für jede Kombination von Visualisierungstechnik und analytischem Modell unterscheidet und daher jeweils neu gelernt werden müsste. Zudem ist auf diese Weise der Grad der Komplexität der transformierten Inhalte nicht beschränkt.
Der Nachteil der automatischen Modellierung besteht darin, dass ein Modell aus sich heraus keine Informationen darüber liefern kann, wie gut es die Phänomene beschreibt. Dieser Nachteil wird im Rahmen dieses Konzeptes dadurch ausgeglichen, dass zwei unabhängige automatische Prozesse indirekt miteinander gekoppelt werden. Das visuelle Feedback koppelt die Transformationen von Mustern zu Modellen und zurück. Beide Transformationen werden so eingesetzt, dass eine jeweils eine Kontrollfunktion über die andere ausüben kann. Beide Transformationen sind komplementär zueinander, und der visuelle Abgleich beruht darauf, dass die Hintereinanderausführung das wahrgenommene Muster reproduzieren soll.
Ein entsprechender Abgleich funktioniert auch auf der Ebene der Modelle, sofern die visuelle Repräsentierung dafür geeignet ist. Über diese Rückkopplung kann untersucht werden,

- ob die automatischen Verfahren im konkreten Fall eine gute Beschreibung der Muster liefern,
- ob die automatischen Verfahren prinzipiell in der Lage sind, eine gute Beschreibung der Muster liefern,
- ob verschiedene Teilmuster durch ein Verfahren unterschiedlich gut beschrieben werden.

Die Rückkopplung erlaubt daher nicht nur eine Bewertung der Ergebnisse, sondern liefert auch Hinweise darüber, warum die Modellierung/Beschreibung eines Musters fehlerhaft sein kann. Der Mensch kann danach entscheiden, ob er durch die Modifizierung der Eingaben - insbesondere der Auswahl der Bereiche, die zum Muster gehören - die Modellierung verbessert, ob er das Verfahren für die Modellierung des Musters verwirft.

Kapitel 4

Realisierung

Im folgenden soll gezeigt werden, wie das Konzept dieser Arbeit mit Techniken für die Analyse multivariater Daten umgesetzt wurde. Die Analyse multivariater Daten ist unter mehreren Aspekten eine Herausforderung:
Hinter der gleichzeitigen Berücksichtigung vieler Attribute bei der explorativen Analyse steht immer die Annahme, dass Abhängigkeiten sich potentiell zwischen beliebig vielen Attributen manifestieren können. Dabei besteht die Möglichkeit, dass eine Beziehung zwischen n Attributen eines Datensatzes erst dann gefunden werden kann, wenn mindestens diese Attribute auch für die Analyse ausgewählt wurden. Man könnte dabei versuchen, immer *alle* verfügbaren Attribute eines Datensatzes in eine explorative Analyse mit aufzunehmen. Dies ist leider weder mit einzelnen automatischen Verfahren, noch mit Visualisierungstechniken in der Praxis möglich.
Die Nutzung automatischer Verfahren für die Dimensionsreduktion wurde in Abschnitt 2.3.5.4 bereits motiviert. Eine der größten Herausforderungen für automatische Verfahren ist dabei der sogenannte *„Curse of Dimensionality“*: Jede Struktur, die sich aus der ungleichen Verteilung der Datenpunkte in einem hochdimensionalen Datenraum ergibt, wird durch die Anzahl der Dimensionen nivelliert.
Eine noch größere Herausforderung ist dieses Problem aus der Sicht der Informationsvisualisierung. Mit *Parallelen Koordinaten* und verschiedenen pixelbasierten Techniken (siehe Abschnitt 2.4.5.3) stehen zwar Techniken zur Verfügung, die potentiell hundert und mehr Dimensionen gleichzeitig darstellen können. Allerdings muss deutlich unterschieden werden, ob eine Visualisierung es erlaubt, viele Dimensionen eines Datensatzes darzustellen, oder ob sie es erlaubt, auch entsprechend hochdimensionale Abhängigkeiten zwischen diesen zu zeigen. In vielen Fällen beschränkt sich die Visualisierung auf die Darstellung mehrerer oder aller Kombinationen von Relationen zwischen wenigen (meist zwei oder drei) Attributen.
Für das Konzept dieser Arbeit soll dediziert eine Technik vorgestellt werden, bei der nicht die Anzahl der Attribute im Vordergrund steht, sondern eine möglichst hohe Komplexität der Relationen dargestellt werden soll. Dieser Anspruch ist mit dem Konzept der Arbeit direkt verbunden: In Abschnitt 3.1.2 wurde dargelegt, dass sich eine Darstellung komplexerer Daten und Muster für den Nutzer nur dann amortisiert, wenn er die Leistung für die Interpretation der Muster nicht selbst erbringen muss. Die vorgestellte Technik unterstützt daher die Interpretationsleistung durch die Integration von Data-Mining-Techniken für die

automatische Modellierung der durch den Menschen gefundenen Muster.
Die letzte Herausforderung der multivariaten explorativen Datenanalyse, die hier erwähnt werden soll, betrifft die Präsentation der Ergebnisse der Analyse. Es handelt sich dabei weniger um ein technisches, als ein methodisches Problem: Ein Ziel der Präsentation ist die Reduktion der Analyse auf die wesentlichen aus ihr folgenden Aussagen. Das Risiko der Präsentation besteht darin, dass Realität auf nur die Aussagen reduziert wird, die überhaupt durch die Darstellung vermittelt werden können. Allein die Tatsache, dass eine scheinbar plausible Beziehung gefunden wird, kann die Möglichkeit präziserer, aber komplexerer Abhängigkeiten maskieren. Im Folgenden wird daher auch gezeigt, wie dieser Herausforderung im Rahmen des vorgestellten Konzepts begegnet werden kann.

4.1 KVMap - Erweitertes Karnaugh-Veitch Diagram

Der im folgenden Abschnitt vorgestellte Prototyp beschreibt eine Kopplung zwischen einer Visualisierungstechnik für hochdimensionale Daten und Methoden der automatischen Datenanalyse im Sinne des Konzepts dieser Arbeit. Das Ziel beim Entwurf dieser Visualisierungstechnik besteht vorrangig darin, möglichst hochdimensionale Abhängigkeiten zwischen einzelnen Attributen eines Datensatzes darzustellen (siehe Abschnitt 4.1.1). Dabei soll einerseits untersucht werden, wo die technischen Grenzen dieser Darstellung liegen, gleichzeitig soll jedoch auch untersucht werden, wo - im Bezug auf die vorgestellte visuelle Abbildung - die Grenzen der Wahrnehmung hochdimensionaler Abhängigkeiten liegen.
Die zweite Komponente dieses Prototyps dient der automatischen Analyse der Interaktion des Nutzers innerhalb der Visualisierung (4.1.4). Die Interaktion erlaubt es dem Nutzer, alle potentiell interessanten Merkmale des visualisierten Datensatzes direkt zu selektieren. Dabei wird vorausgesetzt, dass der Nutzer die selektierten Merkmale als Datenmuster wahrnimmt. Durch die manuelle Selektion definiert der Anwender eine Klassifikation, auf deren Basis eine automatische Modellierung durchgeführt wird.
Das erzeugte Modell wird dem Nutzer auf zwei verschiedenen Arten exponiert. Die Erste ist die Darstellung des erzeugten Modells selbst in Form einer lesbaren Regel (4.1.4.4), die der Anwender selbst als Muster identifiziert hat. Die zweite Variante ist das konfirmative Feedback (4.1.4.3), in dem aus dem automatisch erstellten Modell wieder durch Simulation ein Datensatz erzeugt wird. Für die Analyse erlaubt dies nicht nur den Abgleich zwischen Modell und Muster, sondern auch die Bewertung des automatischen Verfahrens selbst.
Das Ziel bei der Analyse ist die Identifikation von Regeln für die Vorhersage von einer oder mehrerer abhängiger Variablen aus möglichst vielen unabhängigen Variablen eines Datensatzes. Die einzelnen Komponenten des Prototyp haben jeweils einen direkten Bezug zu den verschiedenen Teilen des Konzepts. Die ursprüngliche Motivation des Konzepts ergab sich aus der Untersuchung der Grenzen der menschlichen Wahrnehmung von Mustern. Da die Visualisierungstechnik mehr Variablen zueinander in Bezug setzen kann, als das Arbeitsgedächtnis des Menschen fasst, kommt die Diskrepanz zwischen den Fähigkeiten, Muster zu erkennen und Muster zu interpretieren besonders stark zum tragen.
Eine Visualisierungstechnik, die sich daran orientieren muss, dass alle gefundenen Muster innerhalb der Visualisierung auch interpretierbar sein müssen, wird den tatsächlichen Fähigkeiten des Wahrnehmungssystems also im Allgemeinen nicht gerecht. Die im Konzept vorgeschlagene Aufteilung zwischen Mustererkennung und Musterbeschreibung steht daher vor dem Hintergrund durch eine Spezialisierung der Techniken auf einzelne Teilaufgaben überhaupt die Möglichkeit zu schaffen, dass eine Visualisierung diese Fähigkeiten konsequent ausreizen kann.

4.1.1 Iterative Verfeinerung des Klassifikators

Der Prototyp besteht aus drei Teilkomponenten. Gemeinsam etablieren alle diese Komponenten einen iterativen zyklischen Prozess mit dem Ziel, eine Verbesserung der Klassifikators unter der Kontrolle des Nutzers zu gestalten. Der Anwender kann den Prozess dabei sowohl auf der Ebene der Daten und Musterartefakte, als auch auf der Ebene der Modelle überwa-

chen und steuern. Durch den Abgleich zwischen analytischen Artefakten auf verschiedenen Abstraktionsstufen können jedoch nicht nur die Ergebnisse exponiert werden, sondern auch die Prozesse, die die analytischen Artefakte in beide Richtungen transformieren.
Ein einzelner Zyklus beginnt in der Visualisierung durch die Selektion eines Musters durch den Nutzer, setzt sich fort mit der automatisierten Modellierung des Musters innerhalb der Grenzen des gegebenen analytischen Modells, und endet mit dem Feedback der durch den Klassifikator definierten Mengen innerhalb der Visualisierung.
Die Visualisierungstechnik für die hochdimensionalen Daten und Abhängigkeiten hat drei wichtige Funktionen:

- **Präsentation der Daten:** Aus der Informationsvisualisierung kommt das erste Ziel, dass eine Visualisierung es einem Menschen prinzipiell ermöglichen soll, in den angezeigten Datensätzen verborgene Muster zu erkennen.
- **Interaktive Definition der Klassifikatoren:** Über Interaktionstechniken hinaus, mit denen die Visualisierungstechnik gesteuert werden kann (etwa bezüglich angezeigter Objekte, der Datenattribute oder deren Detaillierungsgrad) soll hier die Interaktionstechniken dediziert auch dafür verwendet werden, die erkannten Muster zu selektieren, und die Information, die zunächst nur als Sinneseindruck vorliegt, in eine Form zu bringen, die der Rechner bearbeiten kann.
- **Intuitive Validierung des Klassifikators:** Da ein Klassifikator Mengen von Objekten beschreibt, liegt der Gedanke nahe, dass ein Klassifikator (oder auch ein anderes deskriptives oder prädiktives analytisches Modell) nur im Kontext dieser Objekte bewertet werden kann. Aus diesem Grund wird hier untersucht, inwiefern die Visualisierungstechnik für die Rohdaten, gleichzeitig auch für die Validierung des Klassifikators eingesetzt werden kann.

Das Klassifikationsverfahren hat zwei Funktionen

- **Beschreibung der durch den Nutzer selektierten Datenmuster:** Die durch den Nutzer identifizierten Muster werden in eine formale Beschreibung umgewandelt.
- **Darstellung des Effekts unterschiedlicher Parametrisierungen:** Dies gilt für Klassifikationsverfahren mit freien Parametern, die unabhängig gesteuert werden können. Wenn ein solcher Parameter Einfluss auf das Modell hat, wird dieser Einfluss im visuellen Feedback dargestellt werden.

Das Simulationsverfahren hat folgende Funktion:

- **Erzeugung eines Datensatzes aus dem Klassifikationsmodell:** Im Prinzip handelt es sich um den umgekehrten Prozess der Analyse. Das Klassifikationsmodell wird auf alle unabhängigen Attribute eines Datensatzes angewendet, um die abhängigen Attribute zu erzeugen. Damit entsteht ein neuer Datensatz, der mit den ursprünglichen Daten verglichen werden kann. Da die Simulation unabhängig von einem bestimmten Analyseverfahren oder Modell ist, kann durch diesen Abgleich die Güte der Modellierung untersucht werden.

Die Modellvisualisierung ist von der Datenvisualisierung zu unterscheiden:

- **Darstellung des Modells in lesbarer Form:** „Lesbar“ bedeutet hier insbesondere eine Form, aus der die gefundene Regel effektiv genutzt werden kann.
- **Manuelle Editierung des Modells:** Bestehende Regeln können verändert werden. Dabei ergibt sich ein direktes Feedback zur Datenvisualisierung über den in der Simulation erzeugten Datensatz. Gleichzeitig können aber auch neue Regeln erzeugt werden. Da auf diese Weise Regeln überprüft werden können, wäre dies der Ausgangspunkt einer konfirmativen Datenanalyse.

Im folgenden Abschnitt werden zunächst die allgemeinen Definitionen und die Formalisierung der Problemstellung beschrieben. Danach werden für diese Realisierung die Komponenten im einzelnen beschrieben.

4.1.2 Formalisierung des Klassifikationsproblems

Insofern Definitionen in mehreren Komponenten des Prototyp gleichermaßen verwendet werden, werden sie in diesem Abschnitt beschrieben. Die Definitionen sollen gleichzeitig auch die Anknüpfungspunkte zwischen den Komponenten beschreiben.
Sei S eine Menge von Datenobjekten. Ein Element $s \in S$ sei definiert als ein Vektor von Attributwerten der Länge dim.

$$s = (s_1, s_2, \ldots, s_{dim}) \tag{4.1}$$

Der Typ der Attribute wird dabei nicht eingeschränkt. Die Attribute sollen dabei im folgenden durch $(A_i)_{i:1..dim}$ bezeichnet werden. A_i bezeichnet dabei gleichzeitig die Menge der Werte des Attributes i.
Für die Definition der Klassifikation sei weiterhin eine Menge $T \subset S$. Diese Menge sei $Targetset$ genannt. Die Klassifikationsaufgabe ist die Identifizierung eines Modells $classifier$, das die Menge T von ihrem Komplement $T^c = S \setminus T$ separiert.
Die Menge T kann beliebig gegeben sein. Im Prinzip wäre eine Auflistung aller ihrer Werte ausreichend. Falls T jedoch interaktiv bestimmt werden soll, ist die manuelle Erstellung einer solchen Liste unpraktikabel. Für die interaktive Definition sei T definiert über eine Menge *abhängiger* Attribute $(A_j^*)_{j \in J}$. Dabei ist $J \subset 1..dim$; es wird also allgemein nicht vorgegeben, welche Attribute des Datensatzes das Targetset beschreiben.
Um T über die abhängigen Attribute zu beschreiben, wird ein einfaches und dennoch allgemeines Modell genutzt, das im wesentlichen jene Elemente zum Targetset zusammenfasst, die gewisse gemeinsame Eigenschaften haben:

$$T = \bigcap_{j=1}^{|J|} \left\{ s : s_j \in Q_j^* \subset A_j^* \right\} \tag{4.2}$$

Dabei bezeichnen die Mengen Q_j^* die *charakteristischen* Werte für die Datenobjekte des Targetset. Natürlich ist diese Beschreibung für T weniger generisch als eine Auflistung beliebiger

Elemente in S, dennoch erlaubt diese Definition über die Wahl der abhängigen Attribute und der charakteristischen Werte eines Attributes eine flexible Beschreibung von großer praktischer Bedeutung. Diese Methode erlaubt es, Bezüge zwischen den verschiedenen Attributen eines Datensatzes zu analysieren.
Das Ziel der Klassifikation ist jedoch nicht die Beschreibung der Menge T durch ein Modell der *abhängigen* Attribute (dies ist ja bereits gegeben), sondern durch ein Modell der *unabhängigen* Attribute. Sei $n = dim - |J|$ die Anzahl der unabhängigen Attribute. Ohne Einschränkung der Allgemeinheit können die unabhängigen Attribute die ersten Einträge des Datenvektors belegen, so dass man die gesuchte Funktion formal so beschreiben kann:

$$classifier\,(s_1, s_2, \ldots, s_n) = \begin{cases} true: & s \in T \\ false: & s \in T^c \end{cases} \tag{4.3}$$

Die Typen der Attribute A_i sind hier nicht vorgegeben. In den Rohdaten können diese Werte nominal, ordinal oder numerisch sein. Dies liegt daran, dass der Prototyp selbst auf einer Kategorisierung aller Attribute arbeitet. Notwendig für die Repräsentierung einer Klassifikation und auch für die Visualisierung der Daten ist daher zunächst die Definition von Kategorisierungen $\Phi_i : A_i \to \mathbb{N}$ auf dem Wertebereich für alle unabhängigen Attribute. Das Ergebnis dieser Funktionen ist ein jeweils Identifikator für die Kategorie, in die ein Datenelement eingeordnet wird.
Jeder Datenvektor wird dabei auf einen Vektor von Kategorien abgebildet. Beispielsweise wird ein Datensatz, der Geschlecht, Alter und Größe und Familienstand eines Menschen enthält, über die vier Kategisierungen in einen Vektor eingeordnet, der nur noch die Kategorien enthält:

$$(\text{'weiblich'}, 40, 1.74, \text{'verheiratet'}) \overset{\Phi_1,\Phi_2,\Phi_3,\Phi_4}{\to} (1, 5, 8, 3) \tag{4.4}$$

$$(\text{'männlich'}, 53, 1.75, \text{'ledig'}) \overset{\Phi_1,\Phi_2,\Phi_3,\Phi_4}{\to} (2, 7, 8, 1) \tag{4.5}$$

Diese mehrdimensionale Kategorisierung wird sowohl die Grundlage für die Visualisierung als auch für die effiziente Berechnung eines Klassifikators.
Das Klassifikationsproblem, wie es hier beschrieben und umgesetzt wird, beschreibt lediglich eine binäre Klassifikation, die eine Menge von ihrem Komplement separiert. Im Prinzip kann diese Beschränkung dadurch aufgehoben werden, dass anstelle eines Targetset eine beliebige Partitionierung von S Grundlage für die zu separierenden Klassen darstellt. Die hier beschriebene Kopplung zwischen visuellen und analytischen Verfahren ist davon unabhängig.
Aus der Definition des Klassifikationsproblems kann man nun konkrete Anforderungen an die einzelnen Komponenten ableiten:
Eine besondere Anforderung, insbesondere für die Visualisierungstechnik, ist die Dimensionalität des Datensatzes. Sowohl der für die Visualisierung zur Verfügung stehende Raum, als auch fundamentale Eigenschaften der menschlichen Wahrnehmung (siehe Abschnitt 2.4.4) setzen enge Grenzen bezüglich der Anzahl von Attributen, die miteinander untersucht werden können. Zwar gibt es heute Visualisierungstechniken wie etwa *Parallele Koordinaten* [Sii00, JFLC08], mit den es prinzipiell möglich ist, beliebig viele Attribute gleichzeitig auf den Achsen darzustellen. Dies ist jedoch nicht gleichzusetzen damit, dass auch beliebige wechselseitige Abhängigkeiten zwischen diesen Attribute sichtbar werden.
Die Definition des Klassifikationsproblems impliziert, dass a-priori nicht klar ist, wie viele

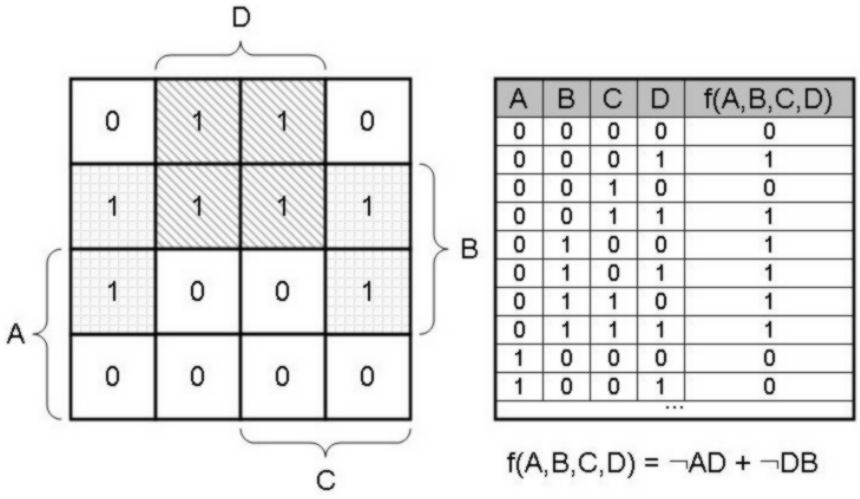
unabhängige Attribute relevant für die Klassifikationsfunktion sind. Ebenso muss nicht bekannt sein, welche Abhängigkeiten zwischen den Attributen herrschen. Das bedeutet, dass im schlechtesten Fall das Weglassen eines einzigen Attributes die entscheidenden Informationen aus dem Prozess entfernt. Ein Muster ist dann, selbst wenn es formal korrekt erkannt und beschrieben wird, nur ein Artefakt das Auswahlprozesses der Attribute.
Daraus ergibt sich einerseits, dass die Anzahl der Dimensionen, die intuitiv zueinander in Bezug gesetzt werden können möglichst hoch sein sollte, und dass die Auswahl der analysierten Attribute des Datensatzes während der Analyse erfolgt. Für die Definition der Partitionierungsfunktionen Φ_i gelten prinzipiell die gleichen Anforderungen. Für die Auswahl dieser Dimensionen gibt es prinzipiell zwei Strategien: Die Auswahl durch den Nutzer und die Auswahl durch eine vorbereitende Analyse: Im ersten Referenzprototyp ist nur die Auswahl der Parameter durch den Nutzer vorgesehen, das dies genügt, um die Durchführbarkeit des Verfeinerungsprozesses prinzipiell zu zeigen.

4.1.3 Modifizierte Karnaugh-Veitch Diagramme

4.1.3.1 Einführung

Ursprünglich wurden Karnaugh-Veitch Diagramme (KV-Diagramme) für die Optimierung Boolescher Ausdrücke bzw. für die Optimierung elektronischer Schaltungen entwickelt [Kar53]. In diesen Diagrammen können prinzipiell alle Booleschen Funktionen mit n Variablen dargestellt werden. Minimale Normalformen für Funktionen können durch das Zusammenfassen von Min- oder Maxtermen erstellt werden. Dabei werden innerhalb des Diagramms Terme identifiziert und vereinfacht, die den gleichen Funktionswert besitzen. Was die Technik von automatischen Verfahren unterscheidet, ist die Tatsache, dass sich ihre Regeln auf *graphische* Muster innerhalb des Diagramms beziehen.
In KV-Diagrammen werden Boolesche Min- und Maxterme nach einem bestimmten Schema auf eine zweidimensionale Karte abgebildet. Für n Funktionsvariablen kann diese Karte dann aus 2^n Zellen bestehen, die alle gemeinsam jede mögliche Kombination von „wahr“ oder „falsch“-Werten der Funktion beschreiben. Jede Variable kann entweder in horizontaler oder vertikaler Richtung ausgelegt werden. Der *Gray Code* aller horizontal angeordneter Attribute definiert die horizontale Position einer Zelle, die zu einer bestimmten Kombination von Variablenwerten gehört. Gleichermaßen gilt das für die vertikal angeordneten Attribute.

In der oben definierten Terminologie stellt die Boolesche Funktion im KV-Diagramm eine Klassifikationsfunktion über n unabhängigen, nominalen Attributen dar, die jeweils zwei Werte annehmen können. Für alle unabhängigen Attribute gilt $A_i = \Phi_i(S) = \{true, false\}$. Die beiden Klassen sind alle Minterme mit dem jeweils gleichen Funktionswert (Null oder Eins). Übersetzt man die Funktionswerte als die Mengen T und T^c wird auch die Aufgabe selbst, die durch die KV-Diagramme gelöst werden soll, also die Suche nach einem optimalen Booleschen Ausdruck für eine gegebene Wahrheitstabelle, zumindest konzeptionell identisch. Man kann die Daten in der Wahrheitstabelle gleichsetzen mit dem Datensatz aus dem Klassifikationsproblem.
Genau genommen ist die Optimierung Boolescher Ausdrücke ein Spezialfall des oben formulierten Klassifikationsproblems, das hier durch mehrere Modifikationen verallgemeinert wer-



A	B	C	D	f(A,B,C,D)
0	0	0	0	0
0	0	0	1	1
0	0	1	0	0
0	0	1	1	1
0	1	0	0	1
0	1	0	1	1
0	1	1	0	1
0	1	1	1	1
1	0	0	0	0
1	0	0	1	0
			...	

Abbildung 4.1: *Eine vierdimensionale Boolesche Funktion dargestellt als Karnaugh-Veitch Diagram (links) und als Wertetabelle (rechts). Die Minimierung der Funktion wird durch die Identifikation von zusammengehörigen, rechteckigen Bereichen mit dem gleichen Funktionswert durchgeführt. Auch die hier vorgestellte Visualisierungstechnik folgt - mit einigen Modifikationen - im Wesentlichen diesem Prinzip.*

den soll. Wichtig und unverändert für die weitere Nutzung bleiben aber zwei Eigenschaften, die es für die Darstellung hoch-dimensionaler Zusammenhänge sehr vorteilhaft sind. Erstens ist die Anzahl möglicher darstellbarer Variablen nicht *prinzipiell* beschränkt und zweitens gehen die Zusammenhänge zwischen einzelnen Variablen auch nach ihrer Abbildung auf zwei Dimensionen nicht verloren.

4.1.3.2 Modifikationen des Layout

Das ursprüngliche Karnaugh-Veitch-Diagramm nimmt die Art der Visualisierungstechnik, die hier vorgestellt werden soll, teilweise vorweg. Es handelt sich dabei um ein Matrix-Layout, das abhängig von der Menge und Größe der einzelnen Zellen mit einem Pixel-basierten Verfahren vergleichbar ist. Darüber hinaus handelt es sich um ein rekursives Matrixlayout, in dem die einzelnen Attribute eines Datensatzes in jeweils einer der beiden Bildschirmachsen ineinander verschachtelt werden. In der Grundkonfiguration (d.h. ohne dass bestimmte Werte gefiltert werden), enthält die Matrix jede Kombination von Werten der dargestellten Attribute genau ein mal.
Schon durch diese Vorgabe ist die Abbildung von höherdimensionalen auf den zweidimensionalen Raum definiert. Um der allgemeineren Beschreibung des Klassifikationsproblems gerecht zu werden, muss das zweidimensionale Layout der KV-Diagramme in mehreren Punkten angepasst werden:

1. Die Anzahl der Partitionen $|\Phi_i(S)|$ für ein Attribut ist nicht notwendigerweise zwei. Prinzipiell ist die Größe einer Partitionierung aus der Sicht der Visualisierung gar nicht festgelegt, sondern kann sowohl durch den Eingriff des Nutzers oder auch durch eine vorbereitende automatische Optimierung der Partitionen verändert werden.

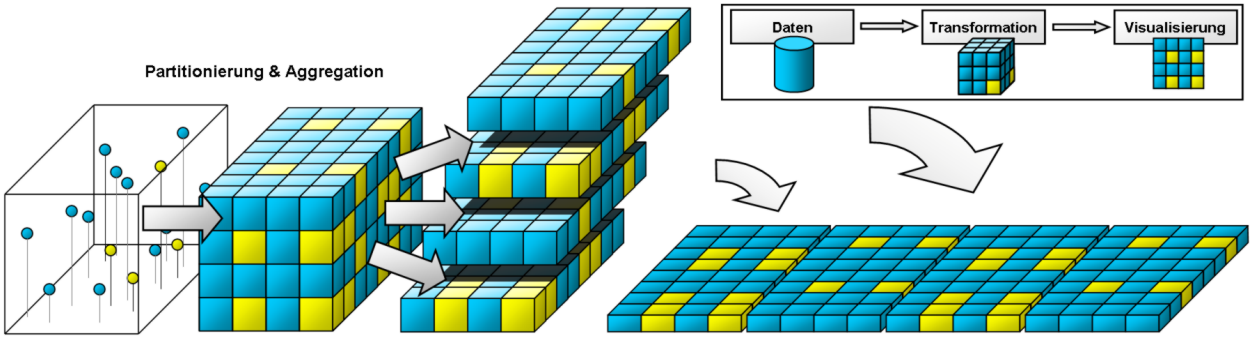


Abbildung 4.2: *Dies ist eine Skizze der visuellen Abbildung des Daten-(Hyper-)-Kubus in dem modifizierten Karnaugh-Veitch Layout. Nach dem Prozess der Datenaggregation(links), wird der Kubus entlang einer Dimension und seiner Partitionierung aufgeteilt. Die Scheiben werden entlang der horizontalen oder vertikalen Bildachse ausgelegt (rechts). Da der Kubus eine Dimension in diesem Prozess verliert, wird er wiederholt, bis alle Dimensionen auf die beiden Bilddimensionen abgebildet sind. Die Information über die räumliche Kohärenz in höheren Dimensionen geht dabei nicht verloren: Die Visualisierung nutzt die Fähigkeit der visuellen Wahrnehmung, verschiedene - auch sich überlagernde - Frequenzen zu identifizieren von denen jede eines der ursprünglichen Attribute repräsentiert. In diesem Beispiel wird das neue Attribut auf die horizontale Frequenz „4“ (d.h. vier Zellen) abgebildet. Zellen in diesem Abstand haben entsprechend ähnliche Werte im bezug auf das dritte, unabhängige Attribut.*

2. Die Position einzelner Zellen im Display wird im modifizierten Layout nicht durch den Gray Code definiert. Teilweise ist dies eine Implikation des ersten Aspekts, da der Gray Code für binäre Werte definiert ist. Mindestens ebenso wichtig ist aber die Tatsache, dass durch den Gray Code einzelne Variablen nicht immer auf unterschiedliche gut visuell abgrenzbare Frequenzen abgebildet werden. Das ist notwendig, da die Frequenzen als Informationsträger eingesetzt werden sollen, die jeweils auch eindeutig einem bestimmten Attribut zuzuordnen sind.

3. Für die Anzeige einer Booleschen Funktion ist es ausreichend, die Zellen mit den Funktionswerten zu markieren. Übertrüge man aber die Klassen T und T^c direkt auf binäre Funktionswerte, würde man gleichzeitig unterstellen, dass die Einträge im Datensatz, die ein bezüglich der Partitionierungen identisches Profil aufweisen auch automatisch entweder *alle* zum Targetset oder *alle* zu dessen Komplement gehören. Davon kann man bei einen beliebig gegebenem Datensatz nicht ausgehen. Selbst eine beliebig feine Auflösung der anzeigten Partitionierungen garantiert noch nicht die Trennung aller Datenelemente in Gruppen, die jeweils ausschließlich der einen oder anderen Klasse angehören. Vielmehr wird für jede Zelle eine statistische Aggregation gebildet, die den Anteil der beiden Klassen für jedes angezeigte Profil als eindimensionale Größe beschreibt. Wie diese Aggregation berechnet wird, wird im folgenden Abschnitt erklärt.

Der erste der genannten Modifikationen besteht darin, dass das Layout der Visualisierungtechnik basierend auf der Wahl der relevanten Attribute und ihrer Partitionierungen angepasst wird. Eine Partitionierung definiert Gemeinsamkeiten und Unterschiede der Datenobjekte, aber es ist a-priori nicht klar, welche Werte für die Klassifikation relevante Informationen enthalten, und wie viele Partitionen gebraucht werden, um diese Information zu beschreiben.

	0	1	2	3	4	...	...	17
0	(0,0,0,0)	(0,0,0,1)	(0,0,0,2)	(0,0,1,0)	(0,0,1,1)	...		(0,0,5,2)
1	(0,1,0,0)	(0,1,0,1)	(0,1,0,2)	(0,1,1,0)	(0,1,1,1)	...		(0,1,5,2)
2	(0,2,0,0)	(0,2,0,1)	(0,2,0,2)	(0,2,1,0)	(0,2,1,1)	...		(0,2,5,2)
3	(0,3,0,0)	(0,3,0,1)	(0,3,0,2)	(0,3,1,0)	(0,3,1,1)	...		(0,0,5,2)
4	(1,0,0,0)	(1,0,0,1)	(1,0,0,2)	(1,0,1,0)	(1,0,1,1)	...		(1,0,5,2)
5	(1,1,0,0)	(1,1,0,1)	(1,1,0,2)	(1,1,1,0)	(1,1,1,1)	...		(1,1,5,2)
...	...	...	...	...	...	...	...	...
...						...		
19	(4,3,0,0)	(4,3,0,1)	(4,3,0,2)	(4,3,1,0)	(4,3,1,1)	...		(4,3,5,2)

Abbildung 4.3: *Das Layout der KVMap ist eine Abbildung der multivariaten Kategorien $(\Phi_1, \Phi_2, \Phi_3, \Phi_4)$ auf die Zellen einer zweidimensionalen Matrix. Diese Tabelle zeigt ein Layout für vier Attribute mit einer unterschiedlichen Anzahl von Kategorien. Die ersten beiden Attribute bestimmen die vertikale, die letzten beiden Attribute die horizontale Position innerhalb der Matrix. Durch das rekursive Layout werden gleiche Kategorien eines Attributes stets im gleichen horizontalen oder vertikalen Abstand dargestellt. Die blauen Zahlen bezeichnen die Zellenpositionen Pos_x bzw. Pos_y (siehe Gleichung 4.7).*

Das eigentliche Layout ist eine verallgemeinerte Variante der b-adischen Darstellung von Zahlen (auch bekannt als *poly-adische* Darstellung). Bis auf weiteres sei zur Vereinfachung der Schreibweise angenommen, dass immer *alle* unabhängigen Attribute auch relevant sind für das Layout des Diagramms. Sei $(h_i)_{i \in 1..n}$ eine Indexmenge, die nur die Attribute beschreibt, die entlang der horizontalen Achsen positioniert werden. Dann beschreibt der Vektor

$$(\phi_{h_1}, \phi_{h_2}, \ldots, \phi_{h_n}) = (\Phi_{h_1}(s_{h_1}), \Phi_{h_2}(s_{h_2}), \ldots, \Phi_{h_n}(s_{h_n})) \tag{4.6}$$

die Partitionen, die das Profil eines Datenobjektes s ausmachen und für das horizontale Layout relevant sind. Die Position der dazugehörigen Zelle ist definiert durch:

$$Pos_x = \phi_{h_n} + |\Phi_{h_n}| \cdot \left(\phi_{h_{n-1}} + \left|\Phi_{h_{n-1}}\right| \cdot \left(\phi_{h_{n-2}} + \ldots |\Phi_{h_2}| \cdot \phi_{h_1}\right)\right) \tag{4.7}$$

Für das vertikale Layout gilt Entsprechendes. Unter der Bedingung, dass Muster, die in verschiedenen Frequenzen auftreten gleich gut wahrgenommen werden können, stehen damit alle angezeigten Attribute und insbesondere auch beliebige Kombinationen ihrer Werte gleichberechtigt gegenüber. Da für alle Attribute gilt $\phi_i < |\Phi_i|$, ist die Funktion invertierbar. Das modifizierte Layout kann also dafür verwendet werden, ein bestimmtes mehrdimensionales Intervall eindeutig zu spezifizieren.
Jede der Zellen ist gleich groß, wobei die Größe durch die Größe des verfügbaren Bildschirmausschnitts und die Gesamtauflösung des modifizierten KV-Diagrams definiert ist. Das hat zur Folge, dass jedem der Attribute eine Frequenz zugeordnet werden kann, die von der Reihenfolge der Attribute abhängt. Dominante Gemeinsamkeiten oder Unterschiede in den angezeigten Daten, die eine Ausprägung in einem der Attribute haben, korrespondieren mit einem entsprechend dominanten Muster mit genau jener Frequenz.
Für die praktische Nutzbarkeit ist dabei der Aspekt von Bedeutung, dass die Reihenfolge der

Attribute nur einen vernachlässigbaren Einfluss auf die Wahrnehmbarkeit der Muster hat. Getestet wurde dies durch künstliche Muster. Testpersonen mußten mehrere, unterschiedlich dominante Muster in verschiedenen Konfigurationen der Visualisierung erkennen. Die Erkennung und Separation dieser Muster war auch dann noch erfolgreich, wenn die Muster sich überlagerten. Nach dem Test gelten jedoch zwei Einschränkungen: Erstens müssen die Testpersonen wissen, wie ein Muster in der KVMap typischerweise aussehen kann. Zweitens kann ein Muster die Visualisierung dominieren, wenn es einen zusammenhängenden Bildbereich ausfüllt; da dies von der Reihenfolge der Attribute abhängt, hebt die Konfiguration unterschiedliche Muster unterschiedlich stark hervor. Dennoch blieben stets alle Muster auch als solche sichtbar. Auf die prinzipielle *Erkennung* von Merkmalen hat die Konfiguration der Visualisierung daher keinen Einfluss.
Dies gilt nicht für das *Verstehen* und noch weniger für die *Orientierung* des Nutzers im Datenraum. Was für Konsequenzen das hat, und welche Lösungsansätze sich aus der Bewertung des Referenzprototyp ableiten lassen, soll am Ende des Abschnitts näher beschrieben werden.

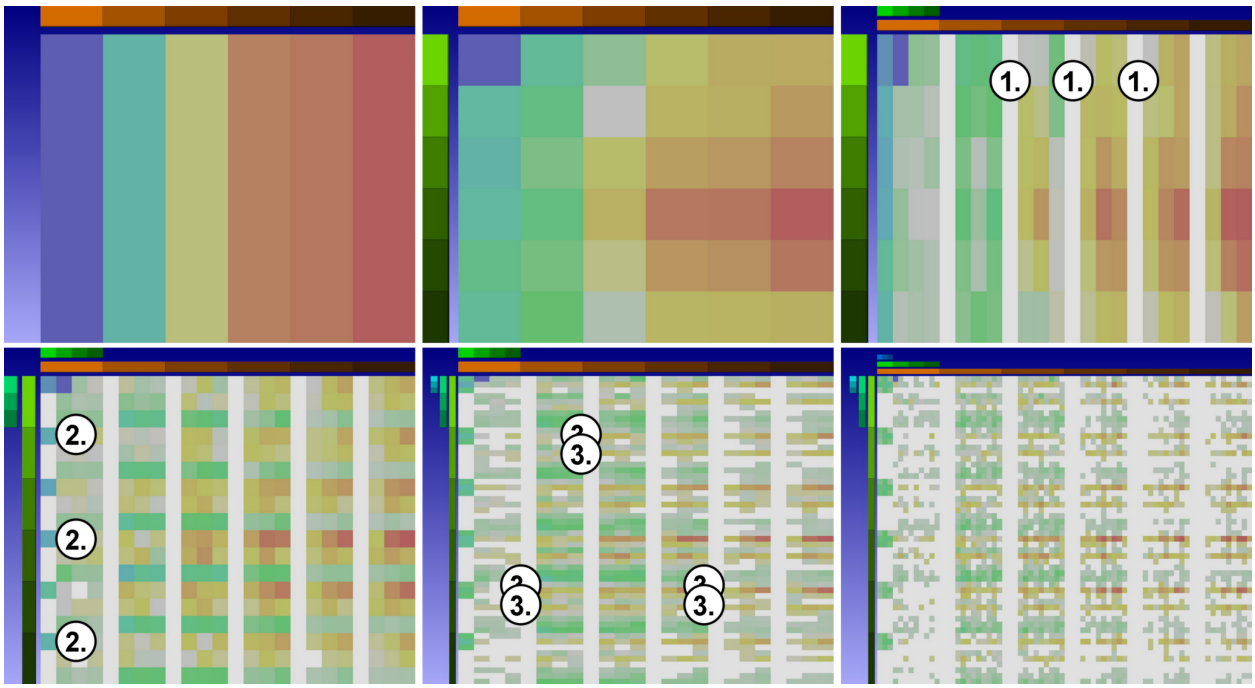


Abbildung 4.4: *Dieses Bild stellt die rekursive Unterteilung der einzelnen Zellen in der Visualisierung dar, die mit jeder Partitionierung eines Attributes feiner wird. Die Farbe beschreibt den statistischen Korrelationskoeffizient, zwischen der Menge, die eine Zelle repräsentiert und dem Targetset. Die dominanten Muster in der Visualisierung sind assoziert mit den dominanten Attributen für die Klassifikation. Die Visualisierung stellt nicht nur die Attribute selbst, sondern auch die Beziehung zwischen diesen Attributen dar. In dieser Darstellung können beispielsweise mindestens drei Muster identifiziert werden, die jeweils von verschiedenen Attributen bestimmt werden.*

4.1.3.3 Statistische Aggregation

Die Modifikationen für das Layout legen nur fest, wie der Datenraum auf den Darstellungsraum abgebildet wird, und dass alle Darstellungselemente für jeweils eine mehrdimensionale Partition in einen rechteckigen Bereich gleicher Größe - die Zellen - passen müssen. Umgekehrt definiert das Layout *nicht*, welche Daten in den Darstellungselementen eigentlich gezeigt werden, und welche visuellen Attribute dafür verwendet werden sollen.
Die Frage nach den verwendeten visuellen Attributen für die Darstellung der Daten ist mehr als eine reine Designentscheidung: Prinzipiell wäre es möglich, Glyphen als Darstellungselemente einzusetzen, die jede für sich wiederum mehrere Attribute darstellen könnte. Allerdings ist die Größe dieser Glyphen begrenzt. Zudem ist es wahrscheinlich, dass die Wahrnehmung komplexer Glyphen mit der Wahrnehmung der Frequenzmuster nicht beliebig kombiniert werden kann [War04b]. Aus diesem Grund wird für das Design zunächst die einfachste denkbare Glyphe gewählt - ein einheitlich gefärbtes Rechteck. Die Farbe des Rechtecks wird durch eine Farbskala definiert, die unabhängig von Layout gewählt werden kann. Die Konsequenz davon ist, dass auch nur ein Aggregat zur gleichen Zeit im KV-Diagramm angezeigt werden kann.
Jede Zelle, identifiziert über ihren Partitionsindex $(\phi_1, \phi_2, \ldots, \phi_n)$ im KV-Diagramm, entspricht einer Teilmenge $P \subset S$, derart dass

$$P(\phi_1, \phi_2, \ldots, \phi_n) = \bigcap_{i=1}^{n} \{s : \Phi_i(s_i) = \phi_i\} \tag{4.8}$$

Man kann nicht davon ausgehen, dass die Menge P einer beliebigen Zelle vollständig zum Targetset oder vollständig nicht zum Targetset gehört. Dennoch sollen die Kenngrößen, die für P berechnet werden, die Beziehung zwischen beiden Mengen beschreiben. Wichtig für die Aufgabe ist beispielsweise die Information, ob die Tatsache, dass ein Element in einer Teilmenge P liegt, einen Indikator dafür darstellt, dass es auch in T oder T^c liegt. Für die Bewertung dieser Informationen stehen mehrere statistische Aggregate zur Verfügung (siehe auch Abbildung 4.5):

- $(|P| \,/\, |S|)$ (Verteilung von S über alle Partitionen)
- $(|P \cap T| \,/\, |S|)$ (relative Mächtigkeit des Targetset)
- $(|P \cap T| \,/\, |T|)$ (Verteilung von T über alle Partitionen)
- $(|P \cap T| \,/\, |P|)$ (Vorhersagekraft von P bezüglich T, dies entspricht der *Konfidenz* der daraus ableitbaren Assoziationsregel)
- $(|P| \cdot |T| - |P \cap T|)) \,/\, \sqrt{|P| \cdot |S \setminus P| \cdot |T| \cdot |S \setminus T|}$ (Der Korrelationskoeffizient nach *Pearson*)

Die Liste ließe sich beliebig fortsetzen. Genau genommen stecken in der Wahl geeigneter Aggregate für die Darstellung der Beziehung zwischen der Verteilung des Targetset und der Verteilung der Objekte des Datensatzes selbst bereits Annahmen darüber, welcher Art diese

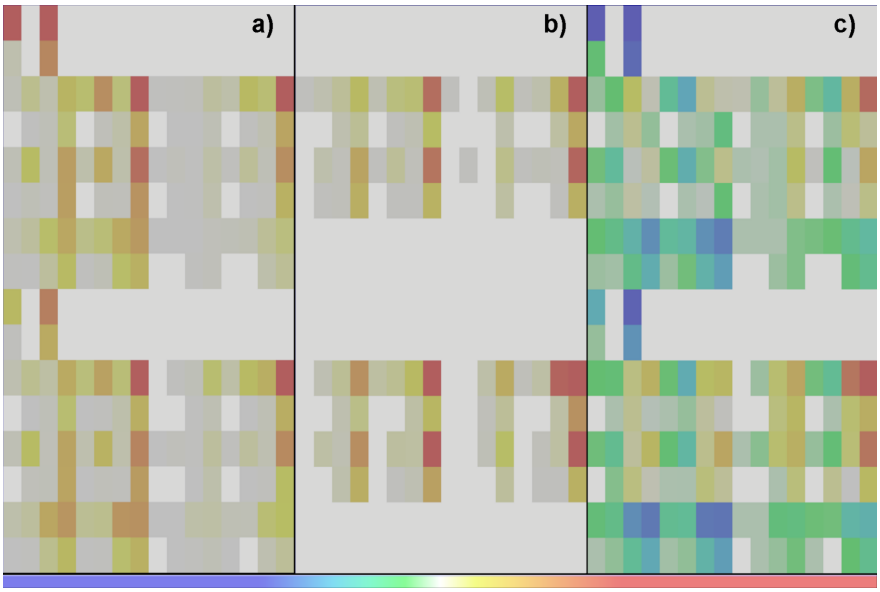


Abbildung 4.5: *Dieses Bild zeigt drei verschiedene Aggregate derselben Daten. Jedes Rechteck repräsentiert eine Partition, d.h. eine Teilmenge der Datenobjekte, die ein spezifisches Profil gemeinsam haben, das durch die unabhängigen Attribute definiert ist. Abbildung (a) zeigt die relative Verteilung aller Objekte S, wobei rot häufige und grau seltene Profile darstellt. Weiße Rechtecke sind leere Partitionen. Abbildung (b) zeigt die relative Verteilung der Objekte in T, die durch die abhängigen Attribute definiert sind. Der statistische Korrelationskoeffizient (c) kombiniert die beiden Werte, wobei rot und gelb positive Korrelation, blau und grün negative Korrelation darstellt.*

Verteilungen tatsächlich sind. Daher sind die verfügbaren statistischen Aggregationsfunktionen ein Verfahrensparameter der Visualisierungstechnik, der frei veränderbar ist. Welche Schlussfolgerungen daraus für die nächsten Prototyp gezogen werden, soll am Ende dieses Abschnittes näher beleuchtet werden.
Drei dieser statistischen Aggregate wurden für die Analyse mit dem modifizierten KV-Diagrammen implementiert: Die Verteilung von T und S über alle Partitionen und der Korrelationskoeffizient für die Korrelation der binären Ereignisse $(s \in P)$ und $(s \in T)$. Die Wahl ist nicht zufällig: Der Korrelationskoeffizient stellt eine Beziehung zwischen den Mengen S und T her, die sich direkt aus den ersten beiden Aggregaten ableiten lässt. Damit kann die komplexere, aber aussagekräftigere Größe auch für den Nutzer sichtbar aus einfacheren Größen hergeleitet werden.
Wie oben beschrieben, wird eine eindimensionale Farbskala verwendet, um die Werte der Aggregation für jede der Partitionen anzuzeigen. Dabei wird bewußt in Kauf genommen, dass Farbe selbst kein geeignetes visuelles Attribut ist, um numerische Werte in relativen und erst recht nicht in absoluten Skalen darzustellen [War04b]. Diese Eigenschaft wird aber auch für diese Visualisierung (bzw. die Aufgabe, die damit durchgeführt werden soll) als nicht relevant angesehen.

Hier sind folgende drei Eigenschaften wichtiger:

1. Eine Farbskala kann unter bestimmten Bedingungen für ordinale Werte genutzt werden.

2. Farbe und Position sind zwei dominierende visuelle Attribute, insbesondere wird Farbe nicht so sehr von der Position des visuellen Elements auf dem Bildschirm dominiert, wie es bei anderen Attributen der Fall wäre.

3. In einem hoch aufgelösten Diagramm kann es sein, dass eine Partition nur wenige Pixel auf dem Bildschirm ausmacht. Farbe funktioniert auch in diesen Fällen.

Brewer [Bre99] definiert Bedingungen, unter denen eine Verwendung von Farbe für ordinale Daten möglich ist. Die erste Eigenschaft ist notwendig, um die Qualität der durch die Aggregation definierten Daten in Beziehung zu einander setzen zu können, um die Frage zu beantworten, bezüglich welcher Tendenz sich *unterschiedlich* gefärbte Partitionen unterscheiden; beispielsweise also, ob eine Partition P eher zum Targetset oder eher nicht zum Targetset gehört.
Die zweite Eigenschaft ist fundamental für die für das Erkennen von Partitionen, die bezüglich der Aggregation identische Merkmale haben, also *gleich* gefärbt sind. Wenn Muster in der Visualisierung sichtbar sind, dann haben sie zwei unabhängige visuelle Attribute als Informationsträger: die Position (bzw. das daraus abgeleitete Merkmal Frequenz) und als zweites Attribut Farbe, die durch die Position nicht oder nur schwach dominiert werden sollte.
Farbe, bzw. die Farbqualitäten Helligkeit und eine Kombination zwischen Farbwert und Sättigung erfüllen die genannten Eigenschaften am besten. Die Wahl fiel letztlich auf den eine eindimensionale Farbwertskala, weil diese zusätzlich noch die Möglichkeit bietet, Farbassoziationen zu nutzen, um „positive“, „negative“ und „neutrale“ Werte darzustellen. Aus diesem Grund ist die Farbskala normiert auf einen Bereich zwischen -1 und 1. Die Farben werden interpoliert von Blau über Grau (für neutrale Werte) nach Rot (für positive Werte). Diese Skala funktioniert sowohl bei der visuellen Trennung von „negativen“ und „positiven“ Aggregationen durch die Trennung von „kalten“ und „warmen“ Farben, als auch bei die Einordnung positiver Werte durch die Zunahme der Sättigung im Bereich zwischen 0 und 1. Die Farben wurden so gewählt, dass die einen möglichst kleinen Helligkeitskontrast (nach dem HSV-Modell *(ebd.)*) aufweisen.
Die Verteilung von leeren Partitionen ist u.U. für die Analyse genauso interessant wie die Verteilung des Targetset. Auf diese Weise können besonders „typische“ Datensätze, Ausreißer, und insbesondere auch Artefakte des Datensatzes (d.h. Kombinationen von Attributwerten, die nicht auftreten können) leicht identifiziert werden. Für die Markierung leerer Partitionen wurde mit mehreren Grautönen experimentiert. Grautöne dunkler als die Farbskala und reines Weiß schieden deshalb aus, weil sie das Bild dominieren können, und so das Auge vom wesentlichen ablenken. Hier wurde ein Grauton gewählt, der gerade so viel heller ist als das „neutrale“ Grau der Farbskala, dass es problemlos davon zu unterscheiden ist, ohne aber das Sichtfeld zu sehr zu dominieren.

4.1.3.4 Interaktion

Nachdem in vorangegangen Abschnitten die Funktionsweise der modifizierten KV-Diagramme beschrieben wurde, soll in diesem Abschnitt zunächst erklärt werden, welche Freiheitsgrade dem Anwender für die Steuerung der Visualisierung zur Verfügung stehen. Anschließend wird erläutert, wie die interaktive Definition der Muster den interaktiven Verfeinerungsprozess für die Erstellung eines Modells auslöst. Die sich daran anschließenden automatischen Prozesse werden in den folgenden Abschnitten beschrieben.
Der erste Freiheitsgrad für die Steuerung ist die Auswahl der unabhängigen Attribute A_i aus der (möglicherweise größeren) Menge der Attribute aus dem Datensatz S. Auch wenn die Visualisierung vergleichsweise viele Attribute gleichzeitig in Bezug setzen kann, so ist diese Anzahl jedoch begrenzt durch

- die Auflösung des Bildschirms
- die Anzahl der Partitionen in jeder Bildachse
- die minimal akzeptable Größe für eine selektierbare Zelle in der Matrix.

Letztere Größe kann abhängig von Zeigegerät und Nutzer variieren. Die Auswahl von Attributen, die potentiell relevante Ergebnisse liefert, ist keine triviale Aufgabe. Faktisch setzt eine „gute“ Auswahl schon viel Wissen voraus, das unter Umständen in der Analyse erst gewonnen werden soll. Der zweite Prototyp setzt sich dediziert mit dieser Problemstellung auseinander (siehe Abschnitt 3.1.4.4). Für diesen ersten Prototyp wurde zunächst eine einfache manuelle Auswahl nach dem Namen eines Attributes umgesetzt. Diese Strategie ist immer dann sinnvoll, wenn die Namen der Attribute eine Bedeutung besitzen, die der Anwender mit seiner Aufgabenstellung verbinden kann.
Wählbar ist jedes Attribut eines Datensatzes. Die KVMap ist insbesondere also nicht abhängig von einem Skalentyp und kann nominale, ordinale und numerische Werte gleichermaßen verarbeiten. Vorraussetzung dafür ist jedoch die Kontrolle über die Partitionierung Φ jedes gewählten Attributes. Auch hier wurde zunächst ein manuelle Steuerung dieser Partitionierung umgesetzt. Die manuelle Steuerung erfolgt dabei über Histogramme (siehe Abbildung 4.6), die den Skalenbereich jedes Attributs abbilden. In jedem Skalenbereich können durch sukzessive Unterteilung beliebige Unterteilungen definiert werden.
Vor der Partitionierung eines Attributs muss jedoch unterschieden werden, ob es sich um ein nominales oder nicht-nominales Attribut handelt. Für ordinale und numerische Attribute ist die Definition von Partitionen über Intervalle effizient und erhält die Topologie der Skala (d.h. ähnliche Werte werden eher in einer Partition zusammengefasst als Unähnliche). In Nominalskalen existiert kein Bezug zwischen den einzelnen Werten. Jedem einzelnen Wert muss eine Partition unabhängig von anderen Werten zugewiesen werden können. Abhängig vom Skalentyp des Attributes, werden die Partitionen daher mit der jeweils passenden Schnittstelle definiert.

Bei beiden Schnittstellen handelt es sich selbst um Visualisierungen, die auch als „Legende“ der KVMap verwendet werden können. Um jedoch als Legende zu dienen, muss eine Korrespondenz hergestellt werden: Dies geschieht in diesem Fall durch *Brushing*. Die Bewegung

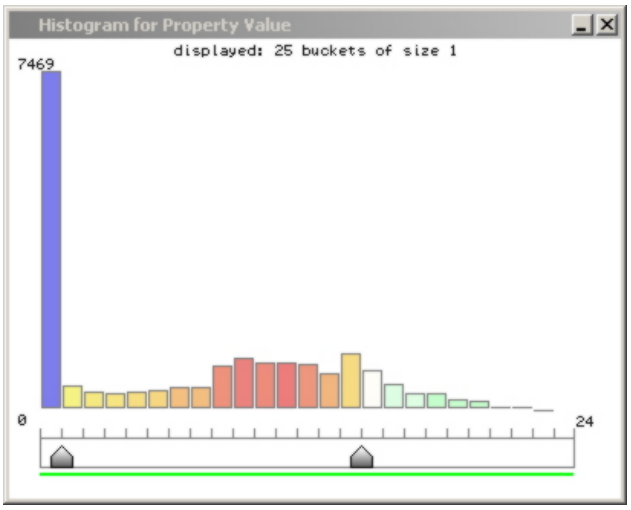


Abbildung 4.6: *Das Histogramm zeigt die Verteilung der Werte eines ordinalen Attributes innerhalb einer Datentabelle. Für die manuelle Steuerung der Partitionierung dieses Attribute kann der Wertebereich in Intervalle unterteilt werden. Die Färbung des Balkens zeigt dabei die Farbcodierung für die gleiche Aggregationsfunktion, die auch für die KVMap verwendet wird.*

des Zeigegeräts über einer Partition bewirkt ein Hervorheben in allen entsprechenden Zellen der KVMap und umgekehrt.
Die KVMap bietet über die gleichen Nutzerschnittstellen auch eine Möglichkeit zum Filtern und Zoomen. Jede einzelne Partition eines Attributes kann in ein Layout ein- oder davon ausgeschlossen werden. Damit können zu jedem Zeitpunkt uninteressante Wertebereiche aus der Darstellung entfernt werden. Dieses entspricht dem Zooming, da das Ausschließen von Partitionen die Möglichkeit bietet, die Anzahl von Zellen in der Darstellung zunächst zu reduzieren. Da die Zellen dabei größer werden schafft das den Platz für eine feinere Unterteilung der übrigen Partitionen oder das Hinzufügen neuer Attribute in das rekursive Layout. Im Prinzip besteht dadurch die Möglichkeit, den Datensatz immer detaillierter darzustellen, und dabei auch alle Attribute eines Datensatzes für einen kleinen Ausschnitt der Daten zu verwenden. Am Ende des folgenden Abschnitts wird eine zusätzliche Möglichkeit vorgestellt, das Filtern und Zoomen semi-automatisch durchzuführen.
In der Praxis ist die Anzahl sinnvoll einsetzbarer Attribute jedoch durch den Umstand begrenzt, dass die Anzahl der Zeilen in einem Datensatz fast immer kleiner ist als die Anzahl möglicher Kombinationen von Partitionen, wenn die Anzahl der verwendeten Attribute genügend groß wird. In diesem Fall bleiben die meisten Zellen leer. Ein Austausch beliebiger Attribute ist dann die geeignetere Strategie für die Exploration des Datensatzes.
Wenn das Targetset T durch die Fragestellung der Analyse abgeleitet werden kann, kann es aus beliebigen externen Quellen definiert werden. Am einfachsten ist es, wenn T durch Werte eines oder mehrere abhängiger Attribute der gleichen Datentabelle definiert werden kann. Zu diesem Zweck werden die gleichen Schnittstellen verwendet, mit denen auch die Partitionen definiert und in die Visualisierung eingefügt oder davon ausgeschlossen werden. Bei der Wahl eines Attributes wird daher zunächst bestimmt, ob es als abhängiges oder unabhängiges Attribut betrachtet werden soll. Auch für abhängige Attribute können beliebige Partitionen definiert werden. Die Partitionen, die *nicht* ausgeschlossen werden, definieren

dann das Targetset gemäß der Gleichung 4.8.
Die im letzten Kapitel 3.1.3.1 vorgestellte wichtigste Form der Interaktion ist die direkte Selektion von Zellen in der KVMap. Die Interpretation der Muster setzt voraus, dass der Anwender genau solche Gruppen von Zellen auswählt, die auch als Teil des desselben Musters wahrgenommen werden. Je mehr Attribute die KVMap darstellt, desto schwieriger wird es für den Anwender, die hinter dem Muster stehende Regelmäßigkeit über den Abgleich mit den einzelnen Partitionen und ihrer Datenwerte durchzuführen. Hier zeigt sich am deutlichsten die Diskrepanz zwischen der Fähigkeit zur Mustererkennung und der Fähigkeit dieses Muster auch effizient zu verstehen und zu interpretieren.
Um ein Muster in der KVMap zu verstehen, müßte der Anwender alle Attribute auf deren Relevanz überprüfen, von diesen wiederum die einzelnen Partitionen, und schließlich müßte er die Teile des Musters identifizieren, in denen sich Abhängigkeiten zwischen zwei oder mehr Attributen manifestieren.
Um diesen Prozess vollständig zu automatisieren, werden die durch den Nutzer identifizieren Zellen als vorgegebene Klassifikation des Datensatzes betrachtet. Ein Modell für diese Klassifikation wird in den folgenden Schritten automatisch erzeugt. Die folgenden Schritte für die automatische Beschreibung eines Musters werden immer dann durchgeführt, sobald der Anwender die Auswahl einer Zelle der Matrix ändert.

4.1.4 Von Musterwahrnehmung zu Modellbeschreibung

Durch den Anwender wurde innerhalb der KVMap eine binäre Klassifikation definiert. Hierbei ist zu betonen, dass die manuell gegebene Klassifikation nicht das Targetset T und sein Komplement $S \backslash T$ beschreibt. Die durch die KVMap dargestellte multivariate Diskretisierung des Merkmalsraums ist im allgemeinen viel zu grob, um das $Targetset$ und sein Komplement vollständig zu separieren. Viele Zellen enthalten daher Datensätze aus beiden Klassen.
Entscheidend ist nicht allein Approximationsgüte, sondern auch die Einfachheit des erstellten Modells. Die Komplexität des Modells wird indirekt dadurch begrenzt, dass sein Muster vom Nutzer noch als Einheit wahrgenommen werden muss. Die vom Nutzer durchgeführte Separierung von Muster und Rauschen, ist bereits der erste Schritt der Analyse. Dabei übernimmt der Betrachter genau jene Aufgabe, der algorithmisch am schwersten zu beschreiben ist (siehe Abschnitt 3.1.1).
Im Folgenden sei $\left(\tilde{T}, S \setminus \tilde{T}\right)$ die durch den Nutzer gegebene Klassifikation. Es wurde für diesen Prototyp eine binäre Klassifikation verwendet. Die folgenden Schritte der Analyse sind ohne Weiteres nicht nur auf binäre, sondern auch auf beliebige weitere Klassifikationen anwendbar. Des Weiteren kann jedes Klassifikationsverfahren eingesetzt werden, das die binäre Klassifikation auf der Basis der Partitionierung der einzelnen Muster durchführen kann. Im vorliegenden Prototyp wurden beispielhaft zwei Verfahren umgesetzt. In beiden Verfahren werden im Prinzip Entscheidungsbäume konstruiert. Die Konstruktion der Entscheidungsbäume unterscheidet sich in beiden Verfahren hinsichtlich der Anzahl ihrer Freiheitsgrade.
Der verwendeten Heuristiken für die Konstruktion der Entscheidungsbäume bestimmen ein Qualitätsmaß für jeden Knoten, anhand dessen in *vollständig* automatischen Verfahren bestimmt werden kann, ab welchen Punkt die rekursive Unterteilung des Baums bei er Konstruktion abgebrochen werden soll. Die in verschiedenen Heuristiken vorgeschlagenen Quali-

tätsmaße haben in diesem Ansatz für die Konstruktion zunächst keine Bedeutung. Die Entscheidungsbäume werden entweder so lange rekursiv aufgebaut, bis entweder eine fehlerfreie Klassifikation in einem Teilbaum erreicht wurde oder bis alle verfügbaren Attribute verwendet wurden. Die Berechnung eines Entscheidungsbaums mit der maximalen Tiefe stellt in diesem Prototyp kein Effizienzproblem dar, da die maximale Anzahl Blattknoten gerade der Menge der Zellen in der KVMap entspricht. Daher können die statistischen Aggregate, die bereits für die Visualisierung berechnet werden, direkt in der Berechnung der Entscheidungsbäume weiterverwendet werden.
Anstatt die Qualitätsmaße für die Konstruktion einzusetzen, dienen sie im Rahmen dieses Konzepts (siehe Abschnitt 3.2.3) als Grundlage für das iterative Feedback. Wie dieses Feedback umgesetzt wird, soll am Ende dieses Abschnitts beschrieben werden. Zunächst sollen die beiden verwendeten Modellklassen vorgestellt werden.

4.1.4.1 Minterme als Klassifikationsmodell

Das erste mit der KVMap gekoppelte Analyseverfahren zeichnet sich dadurch aus, dass es besonders einfach ist, und dennoch viele der Muster, die in der KVMap erkennbar sind, auch durch dieses Modell beschrieben werden können. Der Name leitet sich ab von der Entsprechung für das „Boolesche“ Karnaugh-Veitch-Diagramm. Das Modell repräsentiert die Annahme, dass alle „unabhängigen“ Attribute $(A_i)_{i:1..n}$ tatsächlich auch statistisch unabhängig sind. Die gleiche Annahme wird beispielsweise auch bei der Anwendung des „Naive-Bayes“-Klassifikators getroffen.
Alle Muster, die durch dieses Modell beschrieben werden können, haben stets folgende Form:

$$M = \bigcap_{i=1}^{n} \{s : s_i \in Q_i \subset A_i\} \tag{4.9}$$

Die für alle Attribute A_i zu bestimmendenden Mengen der charakteristischen Werte Q_i sind die freien Parameter des Modells. Die Mengen $Q_i \subset A_i$ sind jedoch nicht beliebig wählbar. Sie sind abhängig von der für die Visualisierung definierten Kategorisierungen Φ_i für jedes Attribut. Es gelte für die Definition der charakteristischen Wertemengen Q_i, dass jeder durch eine Partitionierung $\Phi_i : A_i \to \mathbb{N}$ definierte Wertebereich

$$\Phi_i^{-1}(\phi) = \{v \in A_i : \Phi_i(v) = \phi\} \tag{4.10}$$

immer entweder vollständig in Q_i oder vollständig nicht in Q_i liege. Dadurch wird die Anzahl der Freiheitsgrade des Modells so reduziert, dass die Diskretisierung des KV-Diagramms mit der Definition einer Menge M korrespondiert. M beschreibt in diesem Fall immer eine bestimmte Anzahl von Zellen im KV-Diagramm (bzw. zu ihnen gehörenden Datenobjekte), und jede Zelle liegt entweder vollständig in M oder vollständig nicht in M. Da für jede durch die Interaktion definierbare Menge $\tilde{T}$ das gleiche gilt, ist es nicht sinnvoll, mehr Freiheitsgrade zuzulassen. Die Modellierungsaufgabe ist die Beschreibung einer Menge M, die $\tilde{T}$ möglichst gut approximiert.
Die Einfachheit des analytischen Modells erlaubt eine Separation der einzelnen Attribute. Für jedes der Attribute kann die Menge Q_i unabhängig von der Verteilung der anderen Attribute berechnet werden. Zunächst sei hier daher nur das erste Attribut beschrieben.

Wie im Abschnitt über das Layout beschrieben, ist jede Zelle des KV-Diagramms über ihren Partitionsindex $(\phi_1, \phi_2, \ldots, \phi_n)$ identifizierbar. Nach der interaktiven Auswahl der Zellen kann jedes einzelne Attribut unabhängig untersucht werden. Die Relevanz eines Attributes für das Modell hängt von der Verteilung der Kategorien eines Attributs innerhalb der Auswahl ab. Für jedes Attribut A_i wird daher die Anzahl der ausgewählten Zellen bestimmt, die jeweils zu einer der Kategorien $\Phi_i(S)$ gehören. Mit der Relevanz eines Attributes werden auch die Nummern der Partitionen bestimmt, die für die Menge Q_i in Frage kommen.
Die Idee dabei ist folgende: Wenn die Zellen sich sehr gleichmäßig über die Partitionen eines Attributes verteilen, dann birgt diese Verteilung keine oder wenig Informationen darüber, ob bestimmte Partitionen gegenüber anderen innerhalb des Musters bevorzugt würden. Einzelne Partitionen dieses Attributs haben dementsprechend auch keinen Einfluss auf die Beschreibung der Menge M, d.h. $(Q_i = A_i)$. Ist es umgekehrt so, daß in der Verteilung bestimmte Kategorien eindeutig häufiger auftreten als andere, so ist dies ein Indikator dafür, dass das Attribut relevant ist für die Beschreibung von M. Zwischen diesen beiden Extremen liegt ein weiter Bereich weniger eindeutiger Verteilungen, die letztlich auf eine Auswahl bestimmter Partitionen reduziert werden müssen. Um diesen Bereich zu erfassen, wird der Informationsgehalt einer Verteilung über die bekannte *Shannon-Entropie* definiert.
Sei $h_i(\phi^*)$ die Anzahl der gegebenen Zellen in der Partition ϕ^* des Attributes A_i im Verhältnis zur Anzahl aller Zellen. Die Shannon-Entropie läßt sich aus der normalisierten Verteilung der Partitionen berechnen:

$$E_i = -\frac{1}{log_2 |\Phi_i(A_i)|} \sum_{\phi^*=1}^{|\Phi_i(A_i)|} h_i(\phi^*) \cdot log_2 \left(h_i(\phi^*)\right) \tag{4.11}$$

Diese Gleichung bestimmt die auf das Intervall $[0.0 \ldots 1.0]$ normalisierte Verteilung der Entropie. Je höher dieser Wert ist, desto geringer ist der Informationsgehalt. Die einzelnen Verteilungscharakteristika werden wie folgt zusammengefaßt:

$$E^* \left(\phi_1, \phi_2, \ldots, \phi_n\right) = \prod_{i=1}^{n} \left(h_i \left(\phi_i\right) \left(1 - E_i\right) + E_i\right) \tag{4.12}$$

Diese Funktion verbindet folgende Eigenschaften:

- Partitionen der Attribute werden aufgrund ihrer relativen Häufigkeit für die Menge der charakteristischen Werte selektiert.
- Je höher die Entropie eines Attributes, desto weniger Einfluß hat die relative Häufigkeit auf die Selektion.

E^* ist dabei nicht das Ergebnis des Klassifikators. Es handelt sich um eine Kennzahl, mit der jeder Zelle ein Wert zwischen Null und Eins zugeordnet werden kann. Je höher der Wert, desto eher gehört eine *beliebige* Zelle zu dem Muster, das durch eine Menge $\tilde{T}$ *gegebener* Zellen definiert wurde. Dies ist die Voraussetzung für das visuelle „Fuzzy-"Feedback, wie es in Abschnitt 3.2.3.2 beschrieben wurde.
Die eigentliche Klassifikationsfunktion kann definiert werden mit einem zusätzlichen Thresholdparameter τ:

$$M = \{(\phi_1, \phi_2, \ldots, \phi n) : E^* \left(\phi_1, \phi_2, \ldots, \phi n\right) > \tau\} \tag{4.13}$$

Dieser Thresholdparameter muss zusätzlich zu den selektierten Zellen bestimmt werden werden. Abhängig von diesem Parameter lassen sich dann die Mengen Q_i bestimmen. Die Wahl eines Thresholdparameters bestimmt in vielen automatischen Verfahren die Grenzen zwischen Mustern und Rauschen. Sie ist jedoch kritisch, weil diese Wahl immer vor der Analyse erfolgen muss. Durch die Kombination von visuellen und automatischen Verfahren kann hier dieser Threshold implizit bestimmt werden:
Dafür müssen zwei Voraussetzungen erfüllt sein. Erstens muss die Kennzahl für alle Zellen jedesmal dann neu berechnet werden, wenn sich die Auswahl der Zellen durch den Nutzer ändert. Zweitens muss diese Kennzahl als Feedback der Nutzerinteraktion in der Visualisierung dargestellt werden (siehe übernächster Abschnitt 4.1.4.3). In Abschnitt 3.2.3.1 des vorigen Kapitels wurden die möglichen Ergebnisse des iterativen Feedback vorgestellt. Der Fall, dass das Feedback gegen die Musterwahrnehmung des Menschen konvergiert, schlägt sich in den Kennzahlen nieder. Der Abstand der Kennzahlen für die Zellen, die zum Muster gehören und jenen Zellen, die nicht zum Muster gehören, nimmt dabei kontinuierlich zu.

Die gleichen Modelle lassen sich übrigens auch aus dem *Naive-Bayes* Klassifikationsverfahren konstruieren. Für eine binäre Klassifikation der Menge nach diesem Verfahren gilt folgende Formel, die die *a-posteriori* Wahrscheinlichkeit darstellt:

$$classifier\,(\phi_1, \phi_2, \ldots, \phi_n) = argmax_c p(C = c) \prod_{n}^{i=1} p(\Phi_i = \phi_i | C = c) \tag{4.14}$$

Dabei gilt in diesem Fall $C = \{true, false\}$. Der Wert beschreibt dabei *nicht* die Wahrscheinlichkeiten bezüglich des Targetset T - dies wäre die „normale“ Anwendung des Verfahrens. Stattdessen beschreibt dieser Wert die Wahrscheinlichkeiten bezüglich der Auswahl $\tilde{T}$. Da dieses Verfahren den Klassifikator direkt beschreibt, muss es für das „Fuzzy-“Feedback modifiziert werden. Das „Fuzzy-“Feedback wird insbesondere deshalb notwendig, da ansonsten meist sehr viele Zellen markiert werden müssen, um überhaupt ein Feedback zu erzeugen.
Dass Feedback soll die Tendenz darstellen, ob eine beliebige Zelle das selektierte Muster ergänzt. Anstatt also bei der Klassifikation den wahrscheinlichsten Wert auszuwählen, wird hier die Differenz zwischen den beiden Wahrscheinlichkeiten gewählt. Die Kennzahl ist ein Wert zwischen -1 und 1. Dieser Wert wird für das Feedback visualisiert, das im Abschnitt 4.1.4.3 beschrieben wird.

Durch dieses Verfahren wird ein iterativer Verfeinerungsprozess für die Modellierung eines Musters möglich. Dieses Verfahren wird auf die gleiche Weise auch mit dem nächsten beschriebenen analytischen Modell eingesetzt. Grundsätzlich löst man damit folgendes Problem, dass die Algorithmen, die eingesetzt werden, um die vom Menschen gegebenen Muster zu interpretieren, immer nur Wahrscheinlichkeiten dafür angeben können (und sollen), inwiefern ein bestimmter Teil des Datensatzes gemäß ihres Modells zum beschriebenen Muster gehört. Im Anwendungsfall wird aus diesen Wahrscheinlichkeiten eine (in diesem Fall) „binäre Entscheidung“ durch den Nutzer. Durch den gewählten Kontext kann der Nutzer seine Entscheidung durch den visuellen Abgleich des wahrgenommenen und des berechneten Musters fällen.
Tatsächlich sind die meisten im KV-Diagramm herausstechenden Muster durch dieses einfache Modell zu beschreiben. Der Grund dafür ist, dass Mengen wie sie in der Gleichung definiert sind, durch die dominanten visuellen Attribute dargestellt werden, d.h. durch die

Frequenzmuster. Was geschieht jedoch, wenn die Klassifikation von T und T^c nicht durch eine, sondern durch *mehrere* Mengen $M_1, M_2, \ldots$ erfolgen müsste?
Man kann leicht zeigen, dass jede beliebige Menge - und damit auch das Targetset - sich prinzipiell als Vereinigungsmenge solcher Mengen $M_1, M_2, \ldots$ beschreiben läßt (siehe auch Abbildung 3.23). In der Praxis ist es auch so, dass solche Fälle häufig auftreten. In der Visualisierung bedeutet das, dass sich entsprechend mehrere dieser „einfachen“ Frequenzmuster in einem Bild befinden. Die Konsequenz davon ist aber auch, dass die Minterme für solche Fälle ein bereits zu einfaches analytisches Modell darstellen. Will man *eine* Beschreibung für die gefundenen Merkmale im Datensatz, dann braucht man ein entsprechend komplexeres analytisches Modell, wie es im nächsten Abschnitt vorgestellt wird.

4.1.4.2 Entscheidungsbäume als Klassifikationsmodell

Dieses folgende analytische Modell soll zum einen aus praktischen Überlegungen motiviert und beschrieben werden, die sich aus den Schwächen des vorherigen Modells ergeben haben. Zum anderen aber soll hier untersucht werden, ob es möglich ist, analytische Modelle zu bewerten und zu vergleichen, ohne die Modelle selbst genau kennen zu müssen. Die Frage danach, ob man ein komplexeres Modell einem Einfacheren vorzieht, würde sich nicht stellen, wenn in der Einfachheit kein Vorteil liegen würde. *„Ockham's Razor“* beschreibt im wesentlichen die Forderung nach der Suche des einfachsten Modells, das die wahrnehmbaren Phänomene noch hinreichend genau beschreibt. Über die individuelle Bewertung eines Modells hinaus, muss im Einzelfall auch die Frage diskutierbar werden, ob sich der Aufwand „lohnt“. Das heißt, dass in der Modellanalyse die Frage gestellt werden muß, ob es gerechtfertigt ist, dass mit dem Einsatz eines ausdrucksmächtigeren Modells auch das Risiko eingegangen wird, dass

1. seine Repräsentation komplexer ist,
2. diese Repräsentation sich schwerer in anwendungsbezogenes Wissen umsetzen läßt (dies gilt beispielsweise für viele subsymbolische Modelle, siehe Abschnitt 2.3.3),
3. der Parameterraum größer wird, und Algorithmen für die Parametersuche dementsprechend aufwändiger sind,
4. das Modell so ausdrucksmächtig ist, dass die Gefahr einer Überanpassung besteht, bei der die eingegebenen Daten reproduziert werden, ohne das die berechnete Darstellung einen Informationsgewinn bedeuten würde.

Aus diesen Gründen wird das vorher beschriebene Modell der Minterme dem im Data-Mining ausführlich untersuchten analytischen Modell der *Entscheidungsbäume* gegenübergestellt. Allgemein sind Entscheidungsbäume ein gerichteter azyklischer Graph mit einem ausgezeichneten Wurzelknoten. Der Wurzelknoten ist der einzige Knoten mit $\mathit{In\text{-}Grad} = 0$. Darüber hinaus lassen sich alle Knoten unterteilen in *innere Knoten* (mit $\mathit{Aus\text{-}Grad} > 0$) und *Blattknoten* (mit $\mathit{Aus\text{-}Grad} = 0$). Die mit einem inneren Knoten über ausgehende Kanten verbundenen Knoten heißen *Kindknoten*. Die Baumstruktur modelliert einen sequentiellen,

hierarchischen Ablauf bei der Auswertung des Baumes als Funktion auf einer Menge S.
Die inneren Knoten beschreiben allgemein Partitionierungen auf der Menge S, wobei jedes mögliche Ergebnis mit einem der Kindknoten assoziiert ist. Jede Anwendung des Baums auf ein Element $s \in S$ definiert aus diesem Grund genau einen Pfad, der den Wurzelknoten mit einem der Blattknoten verbindet. Ein Blattknoten beschreibt den Wert der Funktion für die Teilmenge aus S, deren Pfad bei der Anwendung in diesem Blattknoten endet.
Diese allgemeine Definition gibt nur die Struktur des Baumes vor. Sie macht aber weder Annahmen über die Natur der Menge S noch über die konkrete Beschreibung der Partitionierung in den inneren Knoten. Dadurch werden die Entscheidungsbäume zu einem geeigneten Grundgerüst für mehrere Varianten analytischer Modelle, die über die Variation der Komplexität allein der inneren Knoten beschrieben werden können.
Im Vergleich zum vorher beschriebenen Modell sollte das Modell für die Knoten des Entscheidungsbaums daher mindestens so komplex sein, dass der Baum insgesamt die Vereinigung beliebiger Minterme als Muster darstellen kann. Da jede Zelle im KV-Diagramm durch einen Minterm dargestellt werden kann, folgt daraus, dass der Entscheidungsbaum in der Lage ist, beliebige Kombinationen von Zellen und damit eine beliebige Auswahl von Zellen zu repräsentieren.
Jedes Muster, das durch die Minterme darstellbar ist, kann auch über die hier vorgestellten Entscheidungsbäume dargestellt werden. In diesem Sinne sind Entscheidungsbäume eine Erweiterung des einfacheren Modells. Entscheidungsbäume sind darüber hinaus in der Lage, komplexere Muster zu beschreiben. Konkret sind das genau die Muster, die sich aus der Vereinigung beliebiger Minterme ergeben. Über diesen Nutzen hinaus haben Entscheidungsbäume mehrere Vorteile:
Entscheidungsbäume repräsentieren Informationen explizit in symbolischer Form, die durch den Menschen in natürliche Sprache übersetzt werden können (dabei ist die Kenntnis der Bedeutung der Symbole natürlich Vorraussetzung). Damit wird eine formale Beschreibung der Klassifikation mit der Möglichkeit kombiniert, das in der Klassifikation enthaltene Wissen dem Menschen zu exponieren. Im Gegensatz dazu können *subsymbolische* Modelle für die Wissenrepräsentierung (etwa künstliche neuronale Netze) im allgemeinen nicht durch den Menschen „gelesen“ werden.
Entscheidungsbäume bieten ein Grundgerüst für eine ganze Bandbreite analytischer Modelle. Die Komplexität dieses analytischen Modells ist eine Kombination der Komplexität der Baumstruktur und der Komplexität einzelner Knoten des Baums, d.h. der „Entscheidungen“. Die Baumstruktur und der damit eigentlich dargestellte sequentielle, hierarchische Ablauf der Klassifikation ist im Modell fest vorgegeben. Die (inneren) Knoten müssen selbst Klassifikatoren sein, die eine Menge in verschiedene Teilmengen partitioniert.

Entscheidungsbäume können als Modell, wie auch das *Naive-Bayes* Verfahren, nicht direkt für das „Fuzzy“-Feedback eingesetzt werden. Allerdings bieten Entscheidungsbäume von den Daten unabhängige Verfahrensparameter, durch die kontrolliert werden kann, bis zu welcher Baumtiefe der Klassifikator konstruiert oder beschnitten werden soll. Diese „Pruning-Parameter“ (siehe Abschnitt 2.3.5.1) verhindern, dass der Entscheidungsbaum an den Testdatensatz überangepasst wird. Durch diesen Parameter können auf den gleichen Testdaten Entscheidungsbäume konstruiert werden, die unterschiedlich allgemeine Klassifikatoren beschreiben.
Das hier gewählte Verfahren beschreibt Entscheidungsbäume, deren Knoten eine Partitionie-

rung eines einzelnen Attributes in beliebig viele Kindknoten darstellen. Als Partitionierung wird hier stets die die Partitionierung Φ_i verwendet. Die Heuristik, die den Baum konstruiert, muss daher lediglich das Attribut dieses Knotens bestimmen. Dies geschieht über ein Qualitätsmaß, das den Informationsgewinn (*Information Gain*) durch den Knoten beschreibt. In der einfachsten Variante wird dabei jenes Attribut gewählt, das den höchsten Informationsgewinn verspricht.
Der Informationsgewinn wird ausgedrückt über das Entropiemaß, bei dem hier nur zwei Kategorien unterschieden werden; nämlich jene, ob eine Zelle zur Auswahl $\tilde{T}$ gehört oder nicht. Da jeder Knoten des Baums eine Teilmenge $\hat{S} \subseteq S$ beschreibt, läßt sich die Entropie eines Knotens bezüglich der beiden Kategorien definieren wie folgt:

$$H(\tilde{T}, \hat{S}) = -q \cdot log_2(q) - (1-q) \cdot log_2(1-q) \tag{4.15}$$

wobei gilt $q = \left|\tilde{T} \cap \hat{S}\right| / \left|\hat{S}\right|$. Für eine Partitionierung Φ_{attr} eines Attributes, kann die Gesamtentropie über alle Partitionen $P \subseteq S$ - d.h. die Kindknoten - folgendermaßen definiert werden:

$$H_{\Phi_{attr}}(\tilde{T}, \hat{S}) = \sum_{P \in \Phi_{attr}(S)} \frac{|P|}{\left|\hat{S}\right|} H(\tilde{T}, P) \tag{4.16}$$

Der Informationsgewinn eines Attributes und seiner Partitionierung ist dann die Differenz zwischen diesen beiden Größen:

$$Gain_{attr}(\tilde{T}, \hat{S}) = H(\tilde{T}, \hat{S}) - H_{\Phi_{attr}}(\tilde{T}, \hat{S}) \tag{4.17}$$

Der Informationgain ist eine von mehreren möglichen Kennzahlen, mit denen das *Pruning* des Entscheidungsbaums gesteuert werden kann. Die Bestimmung des Wertes für das „Fuzzy"-Feedback funktioniert auch bei anderen Kennzahlen: Der Baum wird dafür bis zur maximalen Tiefe berechnet. Das „Fuzzy"-Feedback ist eine Kenngröße für jede einzelne Zelle der KVMap, die angibt, wie „wahrscheinlich" sie zum bereits selektierten Muster $\tilde{T}$ gehört oder nicht.
Für jeden Blattknoten $\hat{S}$ existiert ein eindeutiger Pfad $S_0 = S, S_2, \ldots, S_n = \hat{S}$. Dieser Pfad beschreibt, zu welchen Knoten der Blattknoten zugeordnet würde, wenn der Baum an den verschiedenen Stufen durch das Pruning beschnitten würden. Entlang dieses Pfads wird der Wert des Klassifikators gewichtet durch den Information Gain für jede Baumsstufe aufsummiert:

$$\sigma\left(\hat{S}\right) = \frac{1}{H\left(\tilde{T}, S\right)} \sum_{i=0}^{n-1} Gain_{attr_i}\left(\tilde{T}, S_i\right) \cdot \left(\frac{\left|S_i \cap \tilde{T}\right|}{|S_i|}\right)^{\alpha} \tag{4.18}$$

Durch die gewichtete Summe wird der Wert des Klassifikators auf jeder Stufe mit der Signifikanz jeder Partitionierung verbunden. $\alpha \in (0..1]$ ist dabei ein Faktor, durch den kontrolliert werden kann, wie viele Zellen des Musters selektiert werden müssen, bevor das Feedback deutlich zu erkennen ist. Kleine Werte für α sorgen dafür dass der Kontrast schon früh zunimmt. Der Ansatz garantiert, dass die Attribute mit dem größten Beitrag zum Klassifikator durch das Feedback auch mit dem deutlichsten Kontrast dargestellt werden.

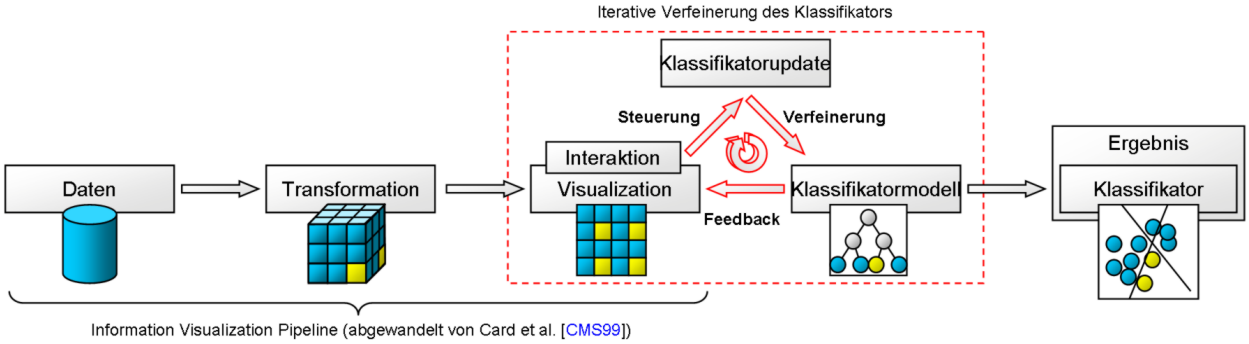


Abbildung 4.7: *Der iterative Verfeinerungsprozess für den Klassifikator interpretiert die Muster aus der Interaktion als Eingabedaten für die abhängige Variable des Klassifikators. Mit jeder Selektion wird das Klassifikator-Modell (ein Entscheidungsbaum) modifiziert. Der Iterationszyklus endet mit der Überblendung des Feedback, das durch den Klassifikator bestimmt wird, über die Originaldaten. Dadurch wird ein Abgleich zwischen dem visuellen/mentalen Modell und dem formalen Modell möglich.*

4.1.4.3 Iteratives Feedback des Klassifikators

Grundlage des iterativen Feedback ist das Qualitätsmaß, das durch die Klassifikatoren jeder Zelle der KVMap zugeordnet werden kann. Durch die automatische Klassifikation werden also insbesondere nicht nur die vom Nutzer direkt selekierten Zellen markiert, sondern auch jene Zellen, die den Rest des selektierten Musters im Sinne des Modells ergänzen würden. Der praktische Vorteil davon ist, das dass der Nutzer nicht alle Zellen eines Musters markieren muss, um schließlich die Musterbeschreibung zu erhalten. Für die Bewertung des verwendeten Verfahrens wichtiger ist jedoch die Möglichkeit, sein Verhalten als Reaktion auf die Nutzereingaben bewerten zu können. Durch das iterative Feedback sollen das Modell, die Menge, die es repräsentiert und das wahrgenommene Muster konvergieren (siehe 3.2.3.1). In dem Prozess beeinflussen sich Nutzer und Verfahren gegenseitig. Wenn Muster und Modell jedoch nicht konvergieren, muss dieses Verfahren als für dieses Muster als ungeeignet erachtet werden.
Für jede Zelle in der KVMap stehen nun drei Daten zur Verfügung:

- die aggregierten Daten (siehe 4.1.3.3)
- die Klassifikation der Zelle durch den Nutzer. Im Falle der binären Klassifikation ist es lediglich ein Flag.
- das Qualitätsmaß des Klassifikators

Die ausgewählten Zellen müssen in jedem Fall leicht zu unterscheiden sein, so dass sie gegebenenfalls wieder deselektiert werden können. Im Prototyp werden sie daher unabhängig von den anderen Daten schwarz gezeichnet. Die aggregierten Daten und das Qualitätsmaß sind Skalarwerte mit bekanntem Maximum und Minimum. Beim iterativen Feedback werden beide Werte auf jede Zelle abgebildet.
Prinzipiell ist ein kein Problem, mehr als einen Wert auf visuelle Attribute abzubilden, was

man zum Beispiel in glyph-basierten Techniken ausnutzt. In diesem Fall sollte jedoch ein visuelles Attribut gesucht werden, das die Erkennung der Muster in den farblich dargestellten aggregierten Daten nicht beeinflusst.
Da die Muster in der KVMap sich als horizontale und vertikale Wiederholungen von Farbeindrücken unterschiedlicher Frequenzen manifestieren, die sich überdies noch überlagern können, kommen bestimmte visuelle Attribute für das Feedback nicht in Frage: Die Veränderung der Position, der Größe oder auch der Form der visuellen Strukturen - d.h. der einzelnen Zellen - kann eine starke Interferenz erzeugen, die als unerwünschtes Artefakt wahrgenommen werden kann. Zusätzlich zu der Farbe werden daher drei andere visuelle Attribute für den Prototyp implementiert und getestet (siehe Abbildung 4.8):

- Helligkeit
- Unschärfe
- Konvexität

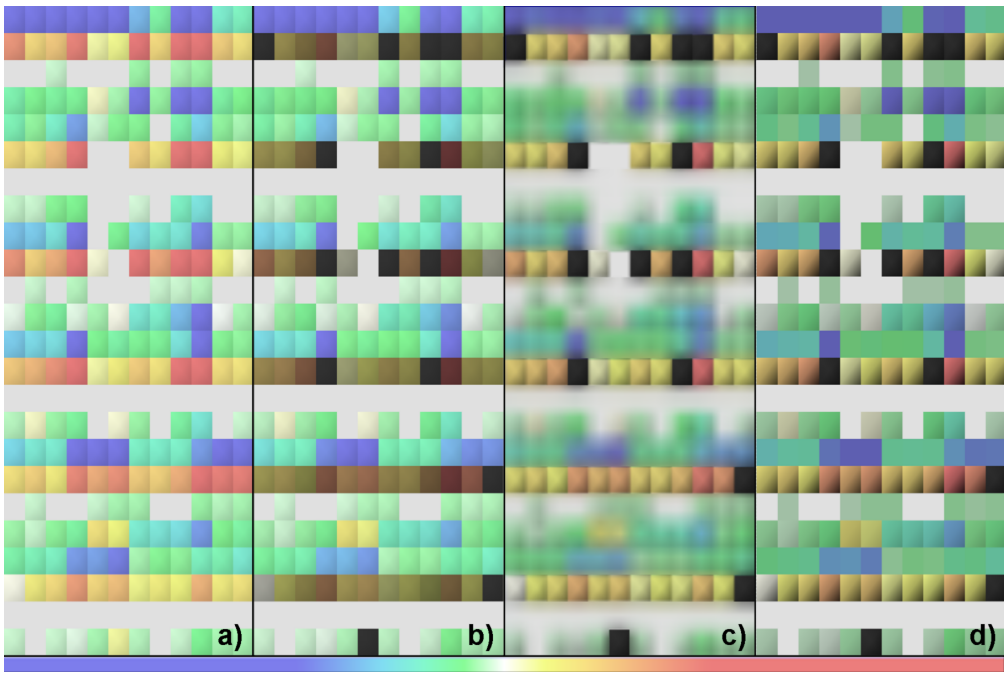


Abbildung 4.8: *Drei visuelle Attribute wurden daraufhin untersucht, unter welchen Umständen die Überblendung mit der Farbe, die die Referenzdaten darstellt (a) funktioniert oder fehlschlägt. Die visuellen Attribute haben unterschiedliche Vor- und Nachteile: Helligkeit (b) ist besser geeignet, wenn viele unabhängige Attribute dargestellt werden und die Rechtecke entsprechend klein sind, aber es dominiert die Farbwahrnehmung am meisten. Unschärfe (c) hat einen kleinen visuellen Wertebereich, d.h. nur wenige Nuancen von Unschärfe können auf einen Blick unterschieden werden. Konvexität (d) stellt einen guten Kompromiss dar, auch wenn eine Mindestgröße der Zellen für die Wahrnehmung dieses Attributes erforderlich ist. Eine Mindestgröße ist aber schon durch die Anforderung der Selektierbarkeit erforderlich.*

Die Auswahl dieser visuellen Attribute basiert auf drei Anforderungen: Alle diese Attribute können nach Ware [War04b] präattentiv wahrgenommen werden. Alle drei Attribute

brauchen keinen zusätzlichen visuellen Raum und die Interferenz mit der Wahrnehmung des Farbtons ist - zumindest bei Unschärfe und Konvexität - vergleichsweise klein.
Die Anpassung des Helligkeitswertes für jede Zelle der KVMap ist direkt umsetzbar, da technisch die Helligkeit gemeinsam mit dem Farbton als Farbwert (RGB) angegeben wird. Im Prototyp werden die Zellen, mit denen der Klassifikator das Muster ergänzt, verdunkelt.
Mit der Konvexität einer Zelle ist die wahrgenommene dreidimensionale Form des Rechtecks gemeint. Der räumliche Tiefeneindruck wird in diesem Fall erzeugt durch einen Schattierungseffekt *(Shading)*, der einen Helligkeitsverlauf darstellt. Dieser wird so gestaltet, dass das Rechteck, entweder als Wölbung in Richtung zum Betrachter oder als Vertiefung wahrgenommen wird. Forsell et al. [FSL05] beschreiben die Verwendung von der Konvexität als visuelles Attribut bei allgemeinen Glyphen. Dabei verwenden sie jedoch Geometrie und ein Modell für die Darstellung von Schatten, um die konvexen bzw. konkaven Formen zu beschreiben.
Die Schattierung ist technisch so implementiert, dass nicht ein einzelnes Rechteck für jede Zelle dargestellt wird, sondern dass stattdessen vier Dreiecke gezeichnet werden, so dass der Mittelpunkt der Zelle ein gemeinsamer Eckpunkt dieser Dreiecke ist. An jedem der fünf Eckpunkte wird dann zusätzlich zum Farbton ein Helligkeitswert zugewiesen.
Das Shading für die Konvexität kann in der KVMap auch anders genutzt werden. In der Matrix kann es geschehen, dass sich benachbarte Zellen im Farbton nur geringfügig unterscheiden. Shading erlaubt dann die visuelle Separation dieser Bereiche, die beispielsweise ohne eine Umrandung der Rechtecke auskommt, die insbesondere bei kleinen Zellen viel Platz beanspruchen würde. Beispielsweise in *Cushion Treemaps* [VWvdW99] wurde diese Technik bereits umgesetzt.
Die Konvexität kann wie auch die Helligkeit ohne große Modifikationen umgesetzt werden. Die Darstellung von Unschärfe allerdings ist aufwändiger in der Umsetzung, da man auf dreidimensionale Graphik zurückgreift. Die Herausforderung besteht darin, unterschiedliche Bereiche des Bildes mit unterschiedlichen Unschärfegraden darzustellen. Dabei wird der Effekt eines Linsensystems emuliert, der genau die Entfernungsbereiche einer Szenerie scharf abbilden, auf die der Betrachter fokussiert. In diesem Prototyp handelt es dabei nicht um Entfernungsbereiche, sondern um die Bereiche der KVMap, die dem Feedback nach eher zum selektierten Muster gehören als der Rest. Der Effekt der Unschärfe für das Hervorheben einzelner Bereiche einer Visualisierung wurde beschrieben von Kosara et al. [KMH01]. Für die Darstellung von Unschärfe wird eine Variante der Technik genutzt, die von Chia et al. [CCAD01] vorgeschlagen wurde. Darin wird die automatische Berechnung von Textur-Mipmaps in der Graphikhardware ausgenutzt. Textur-Mipmaps sind ein Verfahren für die Berechnung verschiedener Auflösungsstufen einer Textur, mit denen Aliasingeffekte bei der Texturierung vermieden werden können. Dabei werden verschiedene Auflösungstufen einer Textur miteinander überblendet. Da in diesem Prototyp keine photorealistische Unschärfe erreicht werden muss, genügt der beim Mipmapping verwendete Boxfilter. Die effiziente Berechnung dieser Stufe ist stattdessen wichtiger, was durch die Nutzung der Grafikhardware erreicht wird.
Im Gegensatz zu den anderen beiden Verfahren bestimmt das Feedback σ keinen Farbwert, sondern die räumliche Tiefe (d.h. die z-Koordinate) in der ein Rechteck der KVMap gezeichnet wird. Für die wahrgenommene Position einer Zelle macht dies keinen Unterschied, da eine Parallelprojektion verwendet wird, in der die Tiefenkoordinate keine Rolle spielt. Allerdings

wird die z-Koordinate des Bildes beim Rendering dazu genutzt, den richtigen Unschärfegrad zu maskieren. Ein Bild der KVMap wird dazu in eine Textur gerendert, die Mipmaps werden automatisch erzeugt und anschließend werden sie in aufsteigender Distanz von Betrachter in die Bildebene gezeichnet. Über die Tiefenkoordinate kann der Renderer effizient die richtige Unschärfe bestimmen.

Die drei verschiedenen visuellen Attribute Helligkeit, Unschärfe und Konvexität wurden auf der Basis folgender Kriterien verglichen:

1. Wechselseitige Beeinflussung mit der präattentiven Wahrnehmung eines Musters in den Originaldaten
2. Qualität der Darstellung in Abhängig von der Auflösung der KVMap (bzw. der Größe der Zellen)
3. Expressivität gradueller Unterschiede

Das visuelle Attribut, das am meisten mit der Identifizierung der Datenmuster wechselwirkt ist die Helligkeit. Der Grund dafür liegt darin, dass die Helligkeit, der Grad der Verdunklung den Kontrast in der Farbskala verringert. Das gleiche Problem würde auftreten, wenn man die Farben der Zellen ausgehend von vollständig gesättigten Farbtönen erhellen würde. Konvexität und Unschärfe dagegen schränken die Wahrnehmung der Datenmuster über die Farbwerte weniger stark ein, was durch die Theorie der Merkmalsintegration von Treisman [TG80] erklärt werden kann: Während Farbton und Helligkeit zwei Farbmerkmale sind, die gemeinsam zu einem Farbeindruck integriert werden, sind die Konvexität und Unschärfe Merkmale für den Tiefeneindruck, der unabhängig davon verarbeitet wird.
Die Größe der Rechtecke in der KVMap, die von der Anzahl der gewählten Partitionen abhängt, beeinflusst die drei visuellen Attribute in unterschiedlichem Maße. Sowohl Unschärfe als auch Konvexität sind visuelle Eindrücke, die sich in einem Bereich des Sichtsfelds manifestieren, anstelle eines einzelnen Bildpunkts. Dieser Bereich muss dafür ein gewisse Mindestgröße haben. Auf sehr kleinen Zellen (20 Pixel und weniger) erzielt man mit Unschärfe, aber auch mit Konvexität keine guten Ergebnisse. Die Wahrnehmung der Helligkeit ist dagegen kaum eingeschränkt.
Die Wahrnehmbarkeit gradueller Unterschiede entlang einer Werteskala ist notwendig, um kleine Änderungen im Wert σ wahrnehmen zu können, wann immer die Selektion der Partitionen sich ändert. Helligkeit und Konvexität können visuell in einem relativ großen Wertebereich separiert werden. Unschärfe hingegen ist ausgesprochen schlecht dafür geeignet, da wenig mehr als drei bis vier verschiedene Grade von Unschärfe im Bild unterschieden werden können.
Die Anzahl der Partitionen, die als Zellen dargestellt werden können, hängt natürlich von der Größe des Displays ab. Allerdings identifiziert Ware [War04b] in einer Arbeit die maximale Größe eines Display (in Pixeln), das abhängig von der Sehschärfe des gesunden menschlichen Auges und des Abstands zum Betrachter noch genutzt werden kann. Die Möglichkeit, prättentiv auch in hochdimensionalen Daten Muster zu erkennen wird in größeren Displays dadurch erschwert, dass der Nutzer den Kopf bewegen und unterschiedliche Sinneseindrucke nacheinander auswerten muss. Die Anzahl und Komplexität der Sinneseindrücke und auch

der zeitliche Rahmen, in dem das geschieht, ist begrenzt. Beliebig große Anzeigegeräte können dieses Skalierbarkeitsproblem nicht lösen.
Auf einem typischen Desktopbildschirm können etwa zehn bis sechzehn Dimensionen gleichzeitig dargestellt werden - abhängig von der Anzahl der Partitionen einzelner Dimensionen. In Experimenten wird die maximal mögliche Auflösung, d.h. ein Pixel für jede Zelle, nie erreicht, da dies die Selektion von Attributen zu sehr behindern würde. Die visuellen Attribute für das Feedback können beliebig gewählt werden, allerdings erscheint die Konvexität in dem meisten Situationen am besten geeignet.

Das Feedback schließt einen Iterationszyklus ab. Jede Selektion oder Deselektion bewirkt eine automatische Veränderung des Klassifikators und damit des Feedback. Die Kohärenz zwischen den beiden Mustern (das der Originaldaten und das des Klassifikators) wird in zwei unabhängigen visuellen Attributen dargestellt. Auf diese Weise kann der Nutzer die Qualität der Approximation eines Musters abschätzen. Bei einer Übereinstimmung beider Muster kann der Anwender den iterativen Prozess beenden und die Repräsentation des Klassifikators in der Form eines Entscheidungsbaums untersuchen.

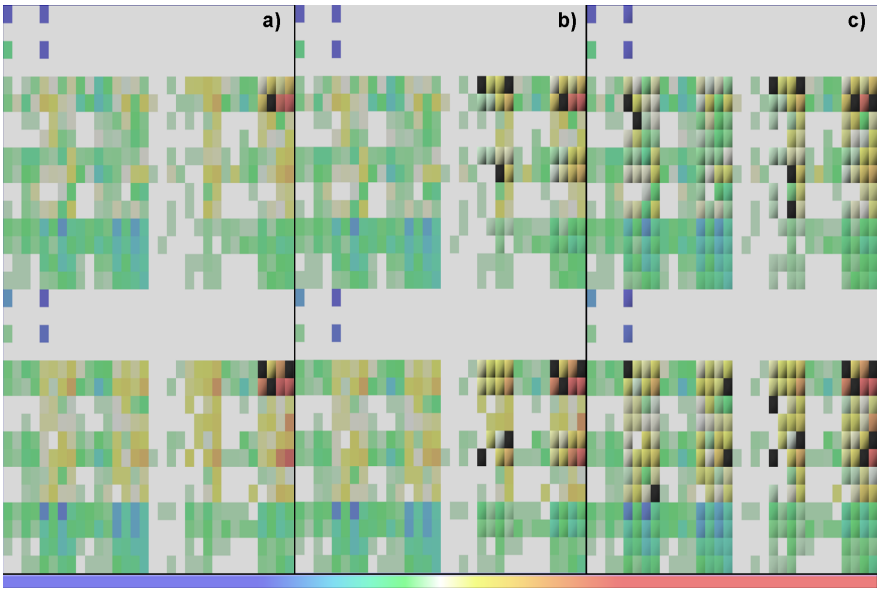


Abbildung 4.9: *Dieses Bild zeigt die Resultate nach 4 (a), 9 (b) und 20 (c) Iterationszyklen für die Verfeinerung des Klassifikators. Jeder Zyklus korrespondiert zu genau einer Partition, die durch den Anwender selektiert (oder deselektiert) wird. Um die Deselektion zu ermöglichen sind diese schwarz. Die Information über die Partitionen, die das visuelle Muster abhängig vom verwendeten Modell ergänzen, werden durch die Konvexitätsattribut dargestellt.*

4.1.4.4 Darstellung des Entscheidungsbaums

Mit der KVMap stellt sich der Unterschied zwischen den Aufgaben *Mustererkennung* und *Musterbeschreibung* in besonders scharfer Form dar. Sie ist so entworfen, dass die Erkennung

von Muster in hochdimensionalen Datenräumen so gut wie möglich unterstützt wird. Das bedeutet auch, dass durch das Design der Visualisierung auch die Grenze dessen überschritten werden, was ein Nutzer noch als „einfache Aussagen“ direkt aus der Visualisierung ableiten kann.
Mit dem Anspruch, mehrdimensionale Daten in einer Form darzustellen, die es erlaubt, auch beliebige, mehrdimensionale Bezüge zu exponieren, nimmt man auch das Risiko in Kauf, dass die entstehenden Muster Aussagen entsprechen, die sich nicht auf eine „griffige“ Formel reduzieren lassen, ohne das dabei viel Information verloren geht. Durch die Kopplung zwischen interaktiver Darstellung und automatischer Analyse kann diesem Risiko begegnet werden. Die KVMap macht die Muster jedoch nur sichtbar, aber nicht lesbar. Der Prozess wird daher ergänzt durch eine interaktive Darstellung des erzeugten Modells.
Für die Darstellung des Entscheidungsbaums ist die Zielsetzung umgekehrt: In erster Linie soll das Muster, bzw. das es beschreibende Modell durch den Nutzer lesbar sein. Erst dadurch können die Informationen einer Interpretation und Evaluierung im Kontext der Problemstellung der Analyse zugänglich gemacht werden. Die beiden Visualisierungstechniken sind durch die automatische Analyse mittelbar gekoppelt. Dies ermöglicht so den parallelen Abgleich zwischen korrespondierenden analytischen Artefakten auf zwei verschiedenen Abstraktionsstufen.
Als Bäume haben die berechneten Klassifikatoren eine visuelle Metapher, die auch in ihrer Darstellung verwendet wird. Das Konzept macht keine Vorgaben darüber, welches Design und welches Layout für die Visualisierung des Baums verwendet werden soll. Beispielhaft, wurde die Visualisierung in enem „Node-Link“-Diagramm umgesetzt wird. In dem gewählten Layout ist der Wurzelknoten stets oben, so dass die Leserichtung für das Modell vorgegeben wird.
Im Entscheidungsbaum werden zum einen Informationen dargestellt, die die Regel beschreiben, zum anderen Informationen, anhand derer die Qualität des Entscheidungsbaums bewertet werden kann. Die Regel wird beschrieben durch die Partitionierung jedes inneren Knotens, und durch die Werte der Klassifikatorregel in jedem Blattknoten des Baums. Die vollständige Partitionierung wird - in Form eines Histogramms - aus Platzgründen nur als Detail-On-Demand angezeigt. Zusätzlich sind in den Knoten folgende Informationen codiert. Die Größe der Oberfläche eines Knotens ist proportional zum Logarithmus der Größe der Menge, die dieser Knoten repräsentiert. Auf diese Weise können die Teile des Klassifikators, die hinsichtlich der relativen Anzahl betroffener Fälle relevanter sind, leichter identifiziert werden. Die logarithmische Skala erlaubt es dabei jedoch, auch noch jene Bereiche des Klassifikators darzustellen, die hinsichtlich ihres Anteils irrelevant sind, aber dennoch abhängig von der Problemstellung der Analyse durchaus die eigentlich interessanten Elemente enthalten könnten. Beispiele dafür sind alle Anwendungsgebiete, in denen die Elemente eines Datensatzes interessieren, die als Abweichung von der Norm untersucht werden müssen.
Die Farbe der Knoten gibt - im Falle einer binären Klassifikation - die relative Anzahl der Elemente eines Knotens an, die zum Targetset gehören $\left(\frac{T \cap S_\nu}{S_\nu}\right)$. Dabei wird eine dreipolige Farbskala verwendet, die die beiden Extremwerte (0 und 100 %) und einen Neutralpunkt enthält. Der Neutralpunkt der Farbskala beschreibt den Anteil, an dem eine Entscheidung darüber, wie die Elemente der Menge binär klassifiziert werden sollen, nicht getroffen werden kann. Dieser Punkt muss nicht zwangsläufig bei 50% liegen. Vielmehr hängt die Wahl des Neutralpunktes auch davon ab, wie viele Fehler 1. und 2. Art man sich im Rahmen der

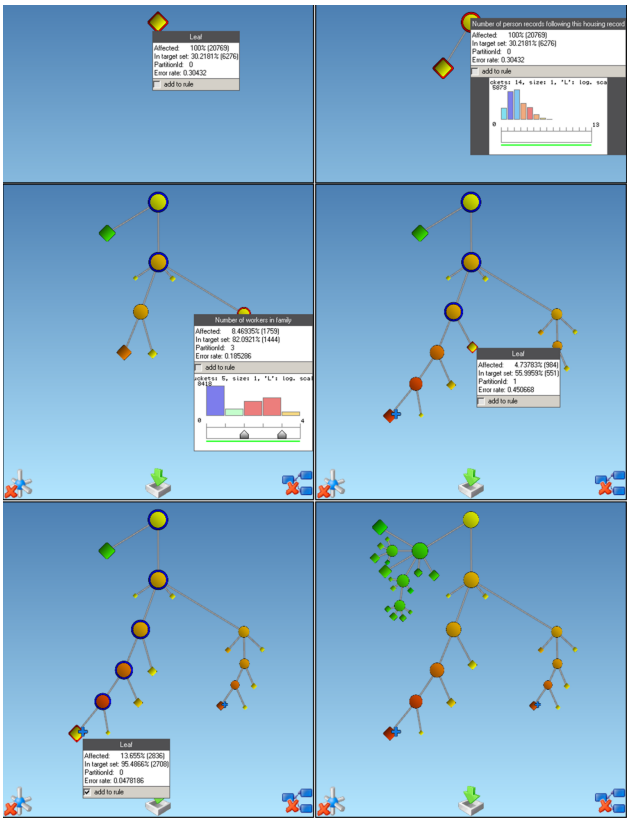


Abbildung 4.10: *Das manuelle Editieren eines Entscheidungsbaums wird hier dargestellt. Beim Editieren kann beispielsweise eine Hypothese als Klassifikatormodell formuliert werden. Ausgehend von Wurzelknoten (links oben), wird für jeden inneren Knoten ein unabhängiges Attribut gewählt, das manuell partitioniert werden kann. Relevanz und Präzision einer „Entscheidung" werden sowohl im Tooltip angezeigt, ist jedoch auch von der Größe und Farbe der Knoten ablesbar. Die Farbe der Knoten repräsentiert die relative Anzahl der Datenobjekte, die zur Zielmenge gehören. Teilbäume, die identisch eingefärbt sind, sind beispielsweise Kandidaten für das Beschneiden des Baums. Bei der Darstellung der automatisch erzeugten Modelle wird die gleiche Visualisierung genutzt.*

Problemstellung der Analyse erlauben darf.
Die Farbskala liefert wertvolle Informationen über die Verteilung der Elemente des Targetset, aber insbesondere auch über die Qualität des Klassifikators. Ein Klassifikator, in dessen Blattknoten nur die Extremwerte vorkommen, erlaubt eine vollständige Separation des Targetset und seines Komplements. Wenn zusätzlich *kein* innerer Knoten existiert, in dem Extremwerte farblich markiert sind, dann ist der Entscheidungsbaum, bezogen auf die Rangfolge der Attribute von Wurzel zu Blattknoten, optimal.

Um die Möglichkeit zu erhalten, berechnete Entscheidungsbäume manuell zu verändern, sowie selbst auch eigene Entscheidungsbäume als Hypothesen in den Analyseprozess einzubringen, ist die visuelle Darstellung des Modells gleichzeitig ein Editor. Dieser Editor ist die Grundlage für die im folgenden Abschnitt beschriebene Simulation. Er erlaubt fünf Grundoperationen auf dem Baum (siehe Abbildung 4.10).

- Hinzufügen eines unabhängigen Attributes zu einem Blattknoten. Dies macht einen Blattknoten zu einem inneren Knoten, der zunächst nur einen Kindknoten besitzt.
- Kopieren eines Teilbaums aus einem inneren Knoten zu einem Blattknoten.
- Löschen eines Teilbaums eines inneren Knotens (manuelles *Pruning*).
- Definition des Klassifikationswertes für einen Knoten. Handels es sich bei dem Knoten um einen inneren Knoten, erhält der darunter liegende Teilbaum den gleichen Klassifikationswert.
- (Re-)Definition der Partitionierung des Attributes für einen inneren Knoten.

Die Klassifikatorregel kann mit diesen Operationen beliebig vereinfacht oder verfeinert werden. Die Partitionierung eines Attributes wird dabei, wie schon bei der KVMap, über ein Tooltipmenu umgesetzt, das ein Histogramm der Werteverteilung für die Elementemenge in diesem Knoten anzeigt.
Die Simulation ist komplementär zur Beschreibung eines Musters in einem analytischen Modell. Die zwei Prozesse - Datenanalyse und Simulation - beschreiben die Transformationen von Musterartefakten zu Modellen und zurück. Da die dabei verwendeten Verfahren unabhängig voneinander sind, erlauben sie den Abgleich zwischen analytischen Artefakten auf zwei Abstraktionsebenen der Analyse.

4.1.5 Simulation mit dem Prädiktiven Modell

Auch wenn „Simulation“ im allgemeinen mit dynamischen, d.h. über die Zeit veränderlichen Prozessen verwendet wird, soll der Begriff hier im allgemeineren Sinne verwendet werden. Allgemeiner beschreibe Simulation hier die Erzeugung neuer Daten, basierend auf einem analytischen Modell unter vorgegebenen Randbedingungen. In diesem Prototyp ist es das gewählte analytische Modell, wobei die Randbedingungen die unabhängigen Attribute des Originaldatensatzes sind.
Das Ergebnis der Simulation ist ein neuer Datensatz, der dem alten Datensatz vollständig gleicht, mit der Ausnahme, dass die abhängigen Attribute der Originaldaten ersetzt werden durch das Ergebnis der Klassifikation. Durch die Klassifikation wird eine neue Menge $T' = \{s \in S : classifier\,(s_1, \ldots, s_n) == true\}$ erzeugt. Die Fragestellung der konformativen Analyse lassen sich im Prinzip auf den Vergleich zwischen beiden Mengen T und T' übertragen.
Ein Grund dafür, eine Klassifikatorregel oder Hypothese durch die Simulation auf den Datenraum zu übertragen, besteht darin, dass die Klassifikatorregel dadurch wiederum beliebigen Visualisierungstechniken aber auch automatischen Verfahren zugänglich wird. Im Normalfall wird bei der konfirmativen Analyse ein Fehlermaß angewendet, um die Differenz zwischen den beiden Mengen zu quantifizieren. Ein Fehlermaß aggregiert jedoch die Daten zu Kennzahlen, aus denen keine Informationen darüber abgeleitet werden können, ob diese Differenz selbst als ein Muster oder als Rauschen qualifiziert werden kann. Stattdessen kann jedoch

jede Visualisierung, die für einen Datensatz S mit einem Targetset $T \subset S$ genutzt werden kann, auch für das Targetset T' verwendet werden. Die beiden Klassifikationen für T und T' können also stets unter identischen Bedingungen einander gegenübergestellt werden.
Technisch betrachtet wird die Klassifikatorregel der Reihe nach auf jedes Element des Datensatzes angewendet. Dabei wird ein neues Attribut erzeugt das - im Falle der binären Klassifikation - angibt, ob der Datensatz zum neuen Targetset T' gehört oder nicht. Mit diesem neuen Attribut kann die KVMap in gleicher Weise eingesetzt werden, wie schon bei den Originaldaten.

4.1.5.1 Visueller Abgleich

Das vorgestellte Konzept soll nicht allein explorative und konfirmative Datenanalyse miteinander verbinden; zusätzlich soll die anspruchsvolle Aufgabe für die Evaluierung einer Hypothese auf einen möglichst präkognitiven Prozess reduziert werden. In diesem Fall ist dieser Prozess der Vergleich zweier Bilder. Die beiden betrachteten Datensätze, die Originaldaten und die simulierten Daten, sind kompatibel zueinander, können also mit der gleichen Visualisierung dargestellt werden. (Die abhängigen Attribute des Originaldatensatzes sind dabei nicht notwendigerweise binär; durch die Definition des Targetset T wird jedoch a-priori eine binäre Klassifikation definiert.)
Zusätzlich muss jedoch sichergestellt werden, dass die Parametrisierung der Visualisierungen für den visuellen Abgleich vollkommen identisch sind. Für die Darstellung der Simulationsdaten wird die Parametrisierung durch das analytische Modell gesteuert. Die wichtigsten Parameter für die Steuerung sind die dargestellten unabhängigen Attribute. Dies gilt im Prinzip für jedes analytische Modell ebenso wie für jede Visualisierungstechnik. Die visuellen Parameter für die KVMap beschreiben die Auswahl der relevantesten Attribute und der Attribute ihres Wertebereichs.
Die unabhängigen Attribute für die Visualisierung werden gewählt nach der Reihenfolge ihres Auftretens im Entscheidungsbaums. Die Rangfolge berücksichtigt dabei die kürzeste Distanz zum Wurzelknoten der Hierarchie, und - in zweiter Priorität - die Anzahl der durch das Attribut klassifizierten Datenelemente.
Bei der Bestimmung der Partitionierung eines Attributes muss berücksichtigt werden, dass eine Partitionierung innerhalb des Baums mehrfach unterschiedlich definiert sein darf. Im Falle der KVMap kann jedoch nur eine Partitionierung verwendet werden. Mehr als eine Partitionierung eines Attributes darzustellen ist dabei kein technisches Problem, sondern behindert die Wahrnehmung der Frequenzen. Die Wahrnehmung von Frequenzen ist die Grundlage für die Erkennung von Mustern in der KVMap. Sie beruht aber darauf, dass jede Frequenz mit genau einen Attribut assoziiert werden kann, was beim Wechsel der Partitionierung nicht mehr garantiert ist.
Um die Visualisierung vergleichbar zu machen, werden die Partitionen $\left(\Phi_i^j\right) : j$ nach folgender Regel vereinigt. Für jeweils ein Paar von Datensätzen s und t gilt:

$$(\Phi_i(s_i) = \Phi_i(t_i)) \Leftrightarrow \bigwedge_{j=1} \left(\Phi_i^j(s_i) = \Phi_i^j(t_i)\right) \tag{4.19}$$

Die so definierte Vereinigungsmenge, ist die gröbste Unterteilung eines Wertebereichs, der alle vorkommenden Partitionierungen berücksichtigt. Jedes Element im Datensatz kann mit einer mehr-dimensionalen Partition $P(s) = (\Phi_1(s_1), \Phi_2(s_2), \ldots, \Phi_n(s_n))$ assoziiert werden. Nach diesem Ansatz entspricht jede Zelle in der KVMap, genau einem Blattknoten im (nicht beschnittenen) Entscheidungsbaum. Es kann also eine direkte Korrespondenz zwischen beiden Visualisierungen hergestellt werden.
Durch die Bestimmung der Vereinigungsmengen der einzelnen Partitionen kann es schnell geschehen, dass die KVMap an die Grenzen ihrer Auflösung kommt. Nur kleine Abweichungen in verschiedenen Definitionen, bewirken eine sehr feine Unterteilung des Wertebereichs. Eine Strategie, um dies zu verhindern, bestünde darin, die entstehende Unterteilung automatisch zu reduzieren. Dies würde jedoch eine der wesentlichen Stärken der KVMap, die Beurteilung der Relevanz eines Attributes, bzw. seiner Unterteilung schwächen. Diese Beurteilung und eine eventuelle Bereinigung der Partitionierung soll weiterhin über die KVMap geschehen.
Stattdessen wird hier die Strategie vorgeschlagen, durch die die Skalierbarkeit der Datenvisualisierung hinsichtlich der Anzahl darstellbarer Attribute über die Visualisierung des Entscheidungsbaums erhöht wird. Der Entscheidungsbaum definiert vom Wurzelknoten abwärts eine Ordnung für die einzelnen Klassifikatoren jedes inneren Knotens gemessen an ihrer Klassifikationsgüte. Da der Baum potentiell eine Tiefe erreichen kann, die die Anzahl der in der KVMap darstellbaren Attribute übersteigt, wird der Baum genutzt, um die Darstellung der Daten in verschiedenen Detailstufen zu steuern. Durch die Wahl eines beliebigen Knoten des Baums kann eine bestimmte Detailstufe für die KVMap definiert werden, die nur die Daten dieses Knotens enthält, und insbesondere die Kategorisierungen aller darüberliegenden Attribute ausblendet.
Zusätzlich zu den in Abschnitt 4.1.3.4 beschriebenen Möglichkeiten für die Darstellung der Datensätze im Detail, lassen sich die gleichen Visualisierungsparameter auch über den Entscheidungsbaum definieren. Der Baum beschreibt zusätzlich die übergeordnete Struktur, über der der Wechsel zwischen Überblick und Detailansichten methodisch gestaltet werden kann.

4.1.5.2 Konfirmative Analyse

Durch die Gegenüberstellung der beiden Datensätze in der gleichen Visualisierung ist es möglich, eine qualitative Bewertung des Klassifikators durchzuführen. Dabei geht es in erster Linie um die Frage, welche Qualität der Fehler des Klassifikators hat. Quantifizierbar ist der Fehler meist über die Bestimmung der Qualitätskennzahlen etwa für die Wahrscheinlichkeit eines Fehlers 1. oder 2. Art. Unabhängig davon jedoch, welche Werte diese Wahrscheinlichkeiten annehmen, muss man berücksichtigen, dass diese Werte selbst immer eine Aggregation über alle Datensätze darstellen. Allerdings verliert man dabei die Information darüber, ob sich dieser Fehler unter unterschiedlichen Bedingungen (d.h. unabhängigen Attributen) unterschiedlichen manifestiert.
Die Kernfrage, die sich bei der Bewertung der Klassifikatoren stellt ist mithin, ob die Verteilung dieser Fehler selbst einem Muster folgt, oder auf ein „Hintergrundrauschen“ der Daten zurückzuführen ist. Hier wird der Standpunkt vertreten, dass eine Bewertung eines Klassifikatormodells über seine Kennzahlen nur dann legitim ist, wenn gleichzeitig bestimmt wurde,

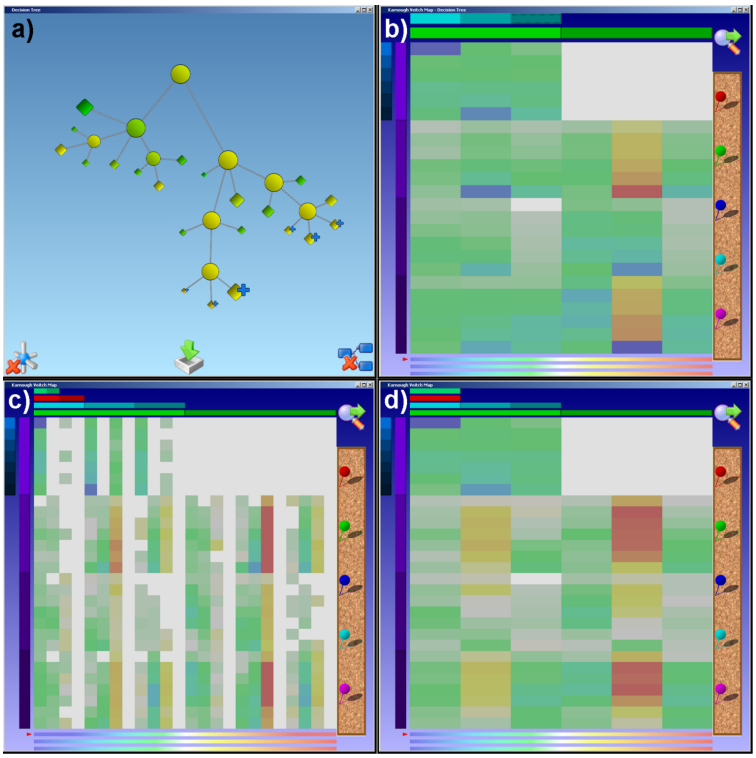


Abbildung 4.11: *Dieses Bild zeigt einen eher schwachen Prädiktor für Haushalte mit einem hohen Grundsteueraufkommen im US-Zensus (a). (b) zeigt eine KVMap mit den vier unabhängigen Attributen des Baums für die simulierten Daten, während (d) die entsprechende KVMap mit den Referenzdaten zeigt. Selbst ohne weitere Informationen, und ohne die Darstellung zu interpretieren, ist ein deutlicher Unterschied sichtbar. Sie zeigt, dass die Hypothese, die in einem Teilbaum modelliert ist, auch in einem anderen Teilbaum anwendbar ist. Eine weitere Analyse (mit mehr Attributen, c), zeigt den Grund für die Schwäche des Klassifikators: Zwei Attribute, die vorher nicht betrachtet wurden (Hauseigentümerstatus und Grundwert) dominieren alle anderen Teile des Musters - sichtbar durch eine dominierende hohe Frequenz - durch die die Farben separiert werden.*

welche Qualität dieser Fehler besitzt. Ein systematischer Fehler ist stets ein Ansatzpunkt für die Verfeinerung des Klassifikators unter Berücksichtigung der vorliegenden Datenbasis. Eine Darstellung der Kennzahlen hingegen verdeckt diese Fehler.

Für die qualitative Bewertung des Klassifikators weniger bedeutsam ist die Untersuchung des Fehlers in einer bestimmten Partition bzw. Zelle in der Darstellung. Vielmehr manifestierten sich systematische Fehler selbst als Muster in den Differenzen zwischen den beiden Bildern. Dazu gehört Beispiel der Fall, dass eine bestimmte Hypothese, die als Entscheidungsbaum formalisiert wurde, nur einen begrenzten Gültigkeitsbereich besitzt. Wenn sich das durch die Hypothese dargestellte Muster auch in den Bereichen der Originaldaten fortsetzt, über die die Hypothese keine Aussagen macht, dann spricht dies dafür, dass sie verallgemeinert werden kann. Die Partitionen, in denen sich das Muster wiederholt können dann in der Entscheidungsregel zusammengefasst werden.
Ein ähnlicher Fall bestünde darin, dass eine an sich korrekte Hypothese auf einer zu schwachen Annahme stützt. Wenn es also ein unabhänges Attribut gibt, das das Targetset trenn-

schärfer beschreibt, dann zeigt die Visualisierung der Originaldaten ein dominanteres Muster, von dem das Muster, das durch die Hypothese definiert ist, lediglich ein Teil dargestellt.
Eine weitere Qualität, die in der Datenvisualisierung untersucht werden kann ist die grundsätzliche Relevanz einer Hypothese. Wenn man den Gültigkeitsbereich einer Hypothese nur weit genug einschränkt ist praktisch immer möglich, in diesem Rahmen zu korrekten Aussagen zu gelangen. Relevant sind die Aussagen jedoch nur dann, wenn sie sich auch dann noch vom Hintergrundrauschen absetzen können, wenn sie im Rahmen eines größeren Zusammenhangs betrachtet werden (z.B. der Gesamtmenge der verfügbaren Daten). Wenn die Hypothese im größeren Zusammenhang nicht mehr vom umgebenden Rauschen unterschieden werden kann, dann ist die Hypothese eher ein Artefakt der Einschränkung ihres Gültigkeitsbereichs, als ein Artefakt der Daten.
Wenn empirische Hypothesen für die Klassifikation von Phänomenen herangezogen werden, dann müssen sie auf der Abstraktionsebene der Daten bewertet werden. Die Visualisierung der Daten stellt einen Kontext bereit für den Gültigkeitsbereich der gerade überprüften Hypothese bereit, der es möglich macht nicht nur die Fehler quantitativ zu beschreiben, sondern auch die Ansatzpunkte für eine Verbesserung zu liefern.
Dabei muss betont werden, dass der vorgestellte Prototyp das Konzept beispielhaft umsetzt. Die Umsetzung ist jedoch nicht beschränkt auf ein bestimmtes Prädiktormodell oder eine bestimmte Analysetechnik. Das Konzept beschreibt eine Methode, mit der verschiedene - unter Umständen auch mehrere - Visualisierungstechniken genutzt werden können, um die Qualität des Klassifikators *und* des Verfahrens, mit dem er erzeugt wurde, qualitativ zu beschreiben.

4.2 Wechselwirkungen zwischen Mustererkennung und Attributselektion

Der vorgestellte Prototyp für die Klassifikation behandelt exemplarisch einen Teilprozess innerhalb der explorativen Analyse. Wie in praktisch allen anderen Techniken für das Data-Mining und der Informationsvisualisierung, hängt die Qualität der gefundenen Muster davon ab, wie gut diese Verfahren gesteuert werden und auf welcher Grundlage die Verfahren gesteuert werden.
Um beispielsweise die KVMap-Visualisierung zu steuern, müssen besonders zwei vorbereitende Aufgaben gelöst werden, durch die die Parameter für den Prototyp bestimmt werden:

- Auswahl der Attribute aus einer hochdimensionalen Datentabelle
- Diskretisierung der Attribute bzw. Partitionierung der Menge der Datenobjekte

Die erste Aufgabe - Auswahl der Attribute - muß für praktisch alle Verfahren gelöst werden, wenn die Datentabelle zu viele potentiell relevante Attribute enthält. Es ist aus mindestens drei Gründen nicht sinnvoll, ein Verfahren zur Mustererkennung direkt auf alle Attribute eines Datensatzes anzuwenden. Erstens ist es häufig technisch gar nicht möglich; dies gilt für Visualisierungstechniken fast immer. Zweitens verwischt die Anzahl der Dimensionen jede potentiell relevante Verteilung des Datenpunkte. Dieses Problem ist als „Curse of Dimensionality“ bekannt. Es liegt darin begründet, dass sich mit zunehmender Anzahl der Dimensionen die Datensätze sich eher am Rand des Hyperwürfels verteilen, der durch die Dimensionen aufgespannt wird. Drittens sind die Attribute in echten Datensätzen praktisch nie unabhängig. Sie enthalten daher meist redundante Informationen. Wenn dies nicht berücksichtigt wird, führt dies dazu, dass verschiedene Abhängigkeiten und Muster in den Daten auf unkontrollierte Weise gewichtet werden. Mit Techniken für die Attributauswahl und - damit verwandt - für die Dimensionsreduktion wird versucht, diese Probleme zu entschärfen.
Die zweite Aufgabe - Diskretisierung der Attribute - ist zwar nicht generisch, dient jedoch ebenfalls als Vorbereitung einer großen Klasse von Verfahren. Die Diskretisierung der Attribute wird durch Binning-Techniken durchgeführt; wobei die Problemstellung allgemeiner als Spezialfall eines Clusteringverfahrens aufgefaßt werden darf. Das Ziel der Diskretisierung ist ja eine Unterteilung der Daten in Teilmengen, die möglichst viel Information über die „objektiven“ Ähnlichkeiten bzw. Unterschiede der einzelnen Datensätze bewahrt.
In Kapitel 2 wurden automatische Verfahren und auch Visualisierungstechniken vorgestellt, die für diese beiden Aufgaben bereits eingesetzt werden. Die beiden Aufgaben sind daher jede für sich ein Forschungsgegenstand eigenen Rechts. Das in diesem Abschnitt vorgestellte Konzept und Verfahren bezieht sich nicht auf eine dieser beiden Aufgaben, sondern auf deren Wechselwirkungen. Die Visualisierungstechnik wurde in den Arbeiten [MDR11] und [MBD+11] bereits vorgestellt.

Nahezu unabhängig davon, wie gut eine dieser Aufgaben einzeln gelöst werden kann, gibt es Wechselwirkungen zwischen der Auswahl der Attribute, der Segmentierung von Daten durch Diskretisierung oder Clusteringverfahren und schließlich auch der Anwendung von Klassifikationsverfahren für die Erstellung eines Prognosemodells. Beispielsweise hängt die Auswahl der Attribute durchaus davon ab, ob man den ganzen Datensatz betrachtet, oder aber nur eine Teilmenge davon. Statistische Abhängigkeiten können sich in unterschiedlichen Teilmengen sehr unterschiedlich ausprägen und damit die Ergebnisse der Auswahl beeinflussen.
Unterschiede zwischen bestimmten Teilmengen werden bei den meisten Verfahren für die Attributauswahl meist nicht berücksichtigt. Zusätzlich bietet eine gute Partitionierung einen Ansatzpunkt für eine Verfeinerung der Analyse, weil es durchaus vorkommen kann, dass verschiedene Verfahren und Modelle angewendet auf unterschiedliche Teilmengen bessere Ergebnisse liefern kann.
Umgekehrt beeinflußt die Attributauswahl die Konstruktion aller multivariaten Partitionen durch Clusteringverfahren oder - indirekt - durch Dimensionsreduktionsverfahren. Für Diskretisierungen eines einzelnen Attributes gilt dies nicht, deswegen werden sie im hier vorgestellten Verfahren als initiale Partitionen genutzt. Univariate Diskretisierungen können jedoch nicht alle potentiell interessanten Teilmengen gut separieren. Zum hier vorgestellten Konzept gehört daher auch eine Strategie, nach der komplexere Partitionen methodisch erzeugt werden.
Warum ist die Berücksichtigung aller dieser Abhängigkeiten in einem einzelnen automatischen Verfahren schwierig? Kriegel et al. [KKZ09] identifizieren *Local Feature Relevance* als eines von vier Problemen, dem mit automatischen Verfahren nur unzureichend begegnet werden kann. Im Prinzip bedeutet es, dass sich verschiedene Cluster (oder auch Muster) in verschiedenen Attributmengen manifestieren, die zudem auch beliebig überlappen können. Auch wenn dieses Problem für Clusteringmethoden definiert wurde, ist es gleichermaßen relevant bei der Klassifikation. Der Suchraum für die Attributauswahl ist die Potenzmenge aller Attribute. Der Suchraum für die Diskretisierung ist die Menge aller Partitionen auf den Daten. Eine Suche auf dem gemeinsamen Suchraum übersteigt derzeit (noch) die Möglichkeiten automatischer Verfahren. Aus diesem Grund soll in diesem Abschnitt ein visuell-interaktiven Verfahren vorgestellt werden, durch das der Mensch in die automatischen Heuristiken eingreifen kann, und das die einzelnen Aufgaben - Attributauswahl, Partitionierung und Mustererkennung - in einem Prozess miteinander verknüpft.

Die zugrundeliegende Idee für das Verfahren wird in der Abbildung 4.12 illustriert. Die Streudiagramme zeigen ein unabhängiges Attribut A_1 und das abhängige Attribut A_t einer Datentabelle. Für die Konstruktion eines Prädiktors ist die Abhängigkeit zwischen A_1 und A_t relevant. Abbildung 4.12a) zeigt die bivariate Verteilung beider Attribute; für kleine Werte wäre A_1 auch tatsächlich ein guter Prädiktor für A_t.
Das Problem besteht nun darin, dass A_1 möglicherweise nur eines von hunderten oder tausenden von möglichen Kandidaten wäre. In so einem Fall ist die Inspektion aller Attribute kaum durchführbar. Aus diesem Grund werden automatische Verfahren für die Attributauswahl genutzt. Fast alle automatischen Verfahren nutzen jedoch *alle* gegebenen Datensätze, um die Qualität eines Kandidaten zu bewerten. Daher würde die starke, lokale Abhängigkeit des Attributes A_t von A_1 wahrscheinlich übersehen. Stattdessen werden schwächere Kandidaten ausgewählt, die im Durchschnitt eine bessere Qualität haben.
Die hier vorgeschlagene Visualisierung stellt die Qualitätsmaße für jedes unabhängige Attri-

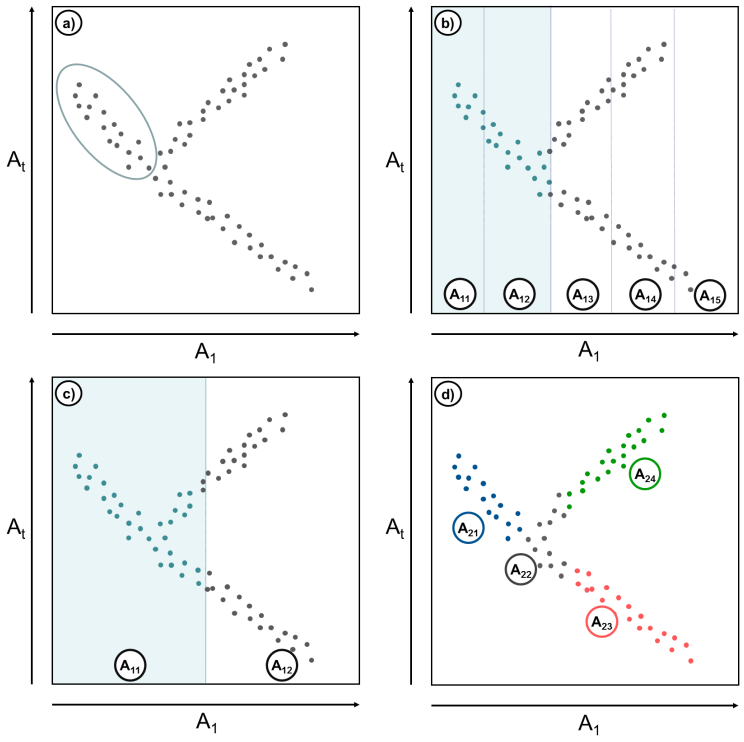


Abbildung 4.12: *Diese Streudiagramme zeigen die zugrundeliegende Idee für die Suche nach Kandidatenattributen A_1 für einen guten Prädiktor für A_t. Die meisten automatischen Verfahren müssen die Maßzahlen für die Abhängigkeit auf der Grundlage aller Datensätze berechnen. Das heißt aber, daß lokale Merkmale von der globalen Verteilung maskiert werden können, selbst wenn sie einen lokal guten Prädiktor liefern würden (a). Um die automatischen Verfahren zu überwachen, wird eine Partition der Daten genutzt, um lokale Details dieser Gütemaße zu untersuchen (b+c+d). Beispiel modifiziert von [KKZ09].*

but visuell dar. Allerdings wird nicht nur die Gesamtqualität untersucht; stattdessen werden die Maße aufgeschlüsselt auf verschiedene Teilmengen des Datensatzes. Abbildung 4.12b) zeigt hierfür eine gute Partition. Bezüglich jeder Teilmenge wird der Beitrag zur Gesamtqualität aufgeschlüsselt und dargestellt. Die lokale, starke Abhängigkeit in den Teilmengen A_{11} und A_{12} hebt sich in der Visualisierung entsprechend ab. Abbildung 4.12c) zeigt eine schwächere Partition des Attributs A_1. Es ist festzustellen, dass die Partition einen großen Einfluß auf die Qualität der Ergebnisse haben. Die Partitionen werden daher nicht festgelegt; sie werden zwar automatisch initialisiert, sind jedoch interaktiv modifizierbar.
Im Prinzip, kann die Aufschlüsselung der Qualitätsmaße sogar entsprechend jeder beliebigen Partition geschehen. Es ist daher auch möglich ein anderes Attribut - anstelle von A_1 oder A_t für die Zerlegung zu nutzen. Abbildung 4.12d) zeigt eine Partition mit einem dritten Attribut A_2, das frei gewählt werden kann. Im folgenden wird gezeigt, wie dieses Verfahren für die interaktive Initialisierung, das Monitoring und die Steuerung eingesetzt werden kann.

Mit der hier vorgestellten Visualisierung werden automatische und visuell-interaktive Attributauswahl verschränkt. Zusammenfassend sollen dabei folgende Anforderungen umgesetzt werden:

- Die Visualisierung soll einen Überblick über eine große, im Prinzip nicht begrenzte, Anzahl von Attributen einer Datentabelle erlauben.
- Die Technik soll auch bezüglich der Anzahl der Datensätze visuell skalierbar sein.
- Die Technik soll auf alle Skalentypen anwendbar sein.
- Sie muß anwendbar sein auf eine Reihe von Verfahren und Qualitätsmaßen für die Attributauswahl. Die Möglichkeit die Qualitätmaße zerlegen zu können sollte die einzige Vorraussetzung für die Anwendbarkeit sein.
- Experten sollten ihr Vorwissen in den Auswahlprozess einbringen können, wobei die automatische Auswahl ggf. durch informierte Eingriffe des Menschen modifiziert werden kann.

Auf der Grundlage existierender automatischer Verfahren und Qualitätsmaße, die im Abschnitt 2.3.5.4 *Dimensionsreduktion* vorgestellt wurden und der obigen Illustration der Problemstellung werden zusätzlich folgende Anforderungen abgeleitet:

- Zerlegung und Visualisierung der Beiträge verschiedener Teilmengen der Daten zum Qualitätsmaß für die Attributauswahl.
- Fokussierung der Attributauswahl auf bestimmte Teilmengen, die durch die Partition eines Attributes definiert werden. Auf diese Weise soll der Attributauswahl an die lokalen Abhängigkeiten bestimmter Teilmengen angepaßt werden.
- Auswahl verschiedener, etablierter Qualitätsmaße und Sortierung der Attribute basierend auf dem gewählten Maß zum Vergleich der Kandidatenattribute.
- Freier Wechsel zwischen einer automatischen und einer interaktiven Heuristik, d.h. die automatische Auswahl kann jederzeit unterbrochen werden, um Attribute von der Auswahl zu entfernen oder hinzuzufügen. Auch die manuelle Initialisierung der automatischen Heuristik wäre so möglich.

4.2.1 Überblick über die Techniken für die Attributauswahl

Der hier vorgestellte Beitrag ist eine Technik für die Auswahl von Attributen für die nachfolgende automatische oder visuell-interaktive Analyse. Es basiert daher auch auf existierenden automatischen Methoden. Wir beschreiben zunächst diese automatischen Techniken. Anschließend werden auch solche Konzepte vorgestellt, in denen die Attributauswahl durch visuell-interaktive Techniken unterstützt wird.
Automatische Methoden für die Attributauswahl müssen für die explorative Analyse eines

Datensatzes eingesetzt werden, die eine große Anzahl von Attributen enthalten. Das Problem kann dargestellt werden als Suche einer geeigneten Teilmenge von Attributen, wobei der Suchraum zunächst alle möglichen Teilmengen von Attributen (d.h. 2^N Teilmengen bei N Attributen) enthält. Selbst nach einer Bereinigung des Datensatzes von bekannten Abhängigkeiten, kann dieser Suchraum immer noch so groß sein, dass eine händische Suche nicht umsetzbar ist. Guyon und Elisseeff [GE03] beschreiben das Problem allgemein als Suche einer *minimalen* Teilmenge an Attributen, die *gemeinsam* für die folgenden Schritte der Analyse den größten Wert haben. Nach ihrer Meinung ist es nicht zu erwarten, dass es eine einzige Technik gibt, die diese Aufgabe in allen möglichen Fällen geeignet löst.
Einen Überblick über die verschiedenen Verfahren für die Attributauswahl und Dimensionsreduktion liefert Abschnitt 2.3.5.4. Das hier vorgestellt Verfahren nutzt dabei die Metriken für die Transinformation (Gleichung 2.1) für die Definition der Abhängigkeit zwischen zwei Attributen und die Systematisierung von Brown (Gleichung 2.2), die garantiert, dass das Verfahren mit sehr unterschiedlichen Metriken verwendbar ist.
Kriegel et al. [KKZ09] schlagen eine Systematik für die Einordnung von Clusterverfahren für hochdimensionale Daten vor. Sie beschreiben die *lokale Featurerelevanz* als eines der vier Probleme, das diese Verfahren lösen müssen. Dies bedeutet, dass unterschiedliche Clusterings sind in unterschiedlichen - nicht notwendig disjunkten - Attributmengen manifestieren. Diese Problemstellung lässt sich ebenso auf Verfahren für die Klassifikation übertragen.

Nachdem Techniken für die automatische Attributauswahl bereits vorgestellt wurden, sollen entsprechende Arbeiten aus der Sicht der Visual Analytics eingeordnet werden. Der hier vorgestellte Ansatz ist vergleichbar mit anderen Methoden, die automatische und visual-interaktive Verfahren kombinieren.
Guo argumentiert in [Guo03], dass die Intervention durch den Menschen notwendig ist, um die Selektion von Attributen zu bewerten und zu steuern. Die Visualisierungen, die am häufigsten für die interaktive Auswahl von Attributen verwendet werden, sind Korrelationsmatrizen. Diese zeigen statistische Testgrößen für Paare von Attributen einer Datentabelle. Die Beziehung zwischen der Verteilung zweier Attributen wird daher auf einen einzelnen Wert kondensiert. Einen Überblick über das Design von Korrelationsmatrizen präsentiert Friendly in [Fri02].
Für eine visuelle Unterstützung der Attributauswahl müssen Information in einem vergleichsweise groben Detailgrad dargestellt werden. Detaillierte Ansichten (d.h. mit allen Werten jedes Datensatzes) sind nicht nutzbar für diese Aufgabe, weil sie fast nie mit der Anzahl der Attribute skalieren. Viele Methoden verfolgen daher eine *Overview & Detail*-Strategie, in der die Übersichtsdarstellung gleichzeitig für die Attributauswahl genutzt wird.
Drei Beispiele, in denen diese Übersicht durch eine Korrelationsmatrix dargestellt wird, präsentieren MacEachren et al. [MDH+03], Seo und Shneiderman [SS04a] und Ingram et al. [IMI+10]. Alle drei Ansätze präsentieren ein Framework, das einen interaktiven Schritt zur Attributauswahl enthält. MacEachren et. el steuern die Attributauswahl direkt über die Matrixdarstellung, wobei das Korrelationsmaß die *maximale bedingte Entropie* ist. Wie oben beschrieben, haben entropiebasierte Testgrößen den Vorteil, dass sie auf alle Skalenniveaus angewendet werden können. Seo und Shneiderman nutzen die Korrelationsmatrizen für die Auswahl interessanter Scatterplots. Für die Beschreibung der Abhängigkeit zweier Variablen können unterschiedliche Maße eingesetzt werden. Ihr *Rank-by-Feature* Framework verbindet zudem die Suche nach guten Clustern auf Grundlage der Auswahl der Attribute. Der hier

vorgestellte Ansatz verfolgt die gleiche Idee und ergänzt sie um die umgekehrte gute Auswahl der Attribute auf Grundlage von Clustern. Ingram et al. fokussieren sich dabei auf die Analyse von Daten mit numerischen Attributen. Die Korrelationsmatrizen dienen der Steuerung von Filteroperationen und der Dimensionsreduktion auf der Grundlage von Pearsons Korrelationskoeffizient.
Die Sortierung der Attribute in der Darstellung ist ein wichtiger Aspekt bei allen visuell-interaktiven Verfahren für die Attributauswahl. Guo [Guo03] schlägt vor, den Wert der Matrixdarstellung durch eine geeignete Anordnung der Attribute zu erhöhen. In seinem Ansatz werden die Attribute nach ihrer Ähnlichkeit sortiert. Die Sortierung im hier vorgestellten Verfahren richtet sich stattdessen nach den Qualitätsmaßen. Die Ordnung, die in einer automatischen Heuristik genutzt werden würde, wird dadurch für den Nutzer sichtbar.
Ansätze, in denen keine Korrelationsmatrix für die Übersicht eingesetzt wird, wurden beispielsweise von Elmqvist et al. [EDF08] und Yang et al. [YHW+07] vorgestellt. Elmqvist et al. nutzen eine Scatterplot-Matrix als Übersicht für die Auswahl von Attributen und für die Navigation in allen achsen-parallelen 2D-Projektionen einer Datentabelle. Scatterplot-Matrizen können eine detaillierte Übersicht über die Verteilung der Daten liefern; der Vorteil geht jedoch mit zunehmender Anzahl der Attribute verloren. Yang et al. schlagen eine Visualisierungstechnik vor, die eine Ansicht der Attributwerte und der Attributbeziehungen kombiniert. Sie zeigt Glyphen, die jeweils uni- oder bivariate Verteilungen darstellen. Das Layout der Glyphen repräsentiert die Ähnlichkeit der Werteverteilungen. Bezüglich der Anzahl der Dimensionen, die gleichzeitig dargestellt und verglichen werden können, ist dies eine der am besten skalierbaren Techniken. Durch die Wahl des Ähnlichkeitsmaßes gilt auch hier die Einschränkung auf numerische Attribute.
Yang et al. präsentieren in [YWRH03] einen Ansatz für die semi-automatische Attributselektion. Ebenso wie der hier vorgestellte Ansatz, ist dieser von einem automatischen Verfahren abgeleitet, das nun interaktiv gesteuert werden kann, um eine brauchbare Teilmenge der Attribute zu bestimmen. Ihr Ansatz berechnet ein agglomeratives, hierarchisches Clustering der Attribute. Für alle Paare von Attributen muß daher ein Ähnlichkeitsmaß bestimmt werden. Numerische Attribute müssen daher kommensurabel sein; ordinale oder nominale Attribute müssen ebenfalls vergleichbar gemacht werden.
Ein weiteres Beispiel für die Integration von automatischen und interaktiven Heuristiken wurde von Ankerst et al. in [AEK00] vorgestellt. Ziel ist die Optimierung eines Entscheidungsbaums. Ähnlich zum hier vorgestellten Ansatz ist die gleichzeitige Optimierung von Attributpartitionen (Bestimmen guter *Split-points*) und die Suche nach relevanten Attributen. Der Unterschied ist, dass unser Ansatz kein analytisches Modell für den Prädiktor vorschreibt.
Johansson et al. [JJ09] stellen einen Ansatz vor, der ein Problem löst, das eng verwandt ist mit dem Problem der *lokalen Featurerelevanz*. Konkurrierende Strukturen in einem Datensatz werden allein durch unterschiedliche Qualitätsmaße hervorgehoben oder aber maskiert. Um dieses Problem zu entschärfen, wählen sie einen interaktiven Ansatz, um verschiedene Qualitätsmaße zu mischen. In dieser Arbeit sind daher die Qualitätsmaße die Freiheitsgrade für die interaktive Analyse.
Piringer et al. [PBH08] stellen ein Konzept vor, mit dem Kandidatenattribute anhand ihrer uni- und bivariaten Verteilungen verglichen werden können. Im Vergleich zu fast allen anderen Ansätzen, werden die Testgrößen nicht nur auf alle Daten, sondern auch auf beliebi-

ge, durch Brushing definierte Teilmengen der Daten angewendet. Unter dem Gesichtspunkt, dass jedes Brushing eine binäre Partition definiert, ist dieser Ansatz dem hier Vorgestellten sehr ähnlich.

4.2.2 Smartstripes - ein Verfahren für semi-automatische Attributselektion

Die Auswahl der Attribute im Rahmen der hier vorgeschlagenen Strategie ist die Bestimmung einer Teilmenge der Attribute, auf die ein analytisches Verfahren in der nächsten Iteration angewendet wird. Im folgenden wird das Verfahren SmartStripes motiviert und beschrieben, mit dem potentielle Abhängigkeiten zwischen beliebigen Attributen unter möglichst geringen Voraussetzungen berechnet, dargestellt und auch gesteuert werden können.
Im folgenden werden die Komponenten dieses Verfahren beschrieben. Wegen der engen Verknüpfung zwischen automatischen und interaktiven Verfahren, sollte der Nutzer die automatischen Verfahren zu einem gewissen Grad verstehen, um damit effektiv umgehen zu können. Zunächst werden wir beschreiben, mit welchen automatischen Verfahren es genutzt werden kann. Da Smartstripes auf einer Diskretisierung der Daten arbeiten muß zeigen wir, wie diese Diskretisierung initialisiert wird, und wie sie in der *Feature Partition View* interaktiv verändert werden kann.
Der Kern von Smartstripes ist die Zerlegung der Qualitätsmaße für die Relevanz und Redundanz (siehe Formel 2.2) in einzelne Teile, um deren Detailgrad für eine Übersicht zu erhöhen. In der *Dependency View* werden diese Details angezeigt und können interaktiv untersucht werden. Beide Ansichten sind miteinander verbunden, und können für die Verfeinerung der automatischen Verfahren - bis hin zu einem Wechsel zwischen automatischer und interaktiver Heuristik - genutzt werden. In Abbildung 4.13 ist ein Schnappschuss des vollständigen Nutzerinterface zu sehen.
Danach werden wir zeigen, wie das Verfahren in dieser Grundform „gelesen“ und genutzt werden kann, welche Stärken und Schwächen Smartstripes in Evaluation und Praxis zeigt. Schließlich beschreiben wir eine Erweiterung des Verfahrens für die iterative Komposition von Analysen und Selektionsverfahren mit Hilfe von synthetischen Attributen.

4.2.2.1 Mit Smartstripes gesteuerte Verfahren

Wie im Abschnitt 4.1.2 *Formalisierung des Klassifikationsproblem* dieses Kapitels beschrieben, sei die zugrundeliegende Aufgabe ein Klassifikationsproblem innerhalb einer Menge von Datenobjekten, die in Form von mehrdimensionalen Vektoren beschrieben werden können. Eine der Voraussetzungen, die das Verfahren erfüllen soll, dass die Vektoren Attribute beliebigen Skalentyps enthalten müssen. Weiterhin gilt die Annahme, dass die Datenobjekte unabhängige Entitäten darstellen. Eine Umordnung von Datenobjekten verändert also nicht die Bedeutung der Daten.
Die Attributauswahl für die Klassifikation ist eine Problemstellung der Skalierbarkeit von Techniken und Methoden, die sich mit wachsender Anzahl von Attributen verschärft. Dies betrifft einerseits die technische Skalierbarkeit von Klassifikationsverfahren, denn die Anzahl

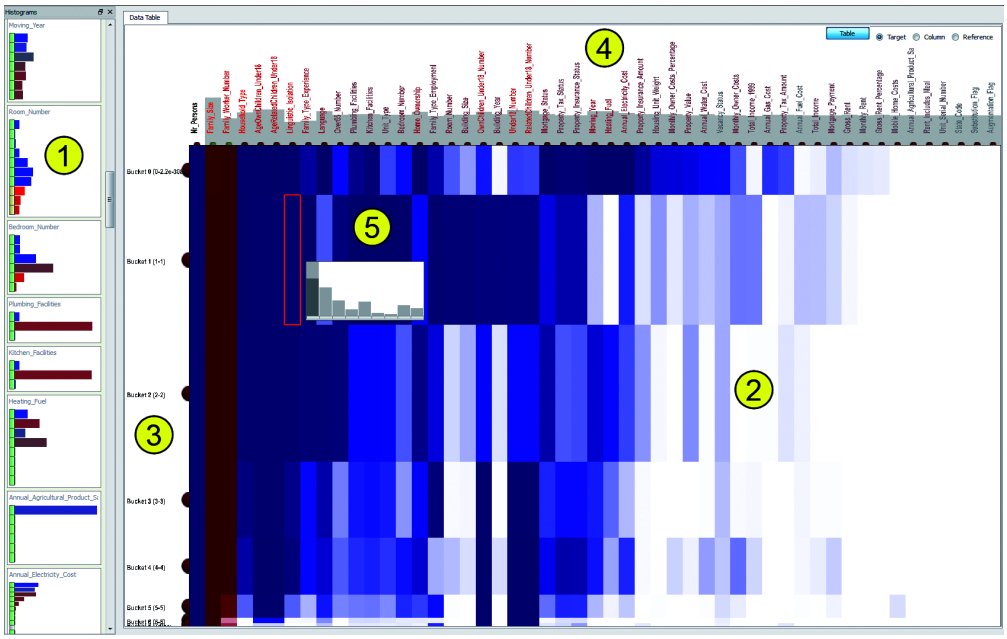


Abbildung 4.13: *Das Bild gibt einen Überblick über das Smartstripes-Verfahren. (1) Feature Partition View. (2) Dependency View. (3) Labels für die einzelnen Partitionen des Referenzattributes. (4) Labels für die einzelnen Attribute. Die Relevanzmaße werden in einem Balkendiagramm dargestellt. Das Targetattribut steht auf der linken Seite (schwarzes Label), die Sättigung beschreibt die lokale Abhängigkeit eines Attributes vom Referenzattributs (siehe Text). Um die Abhängigkeit erklären zu können, liefert ein Tooltip (5) den Vergleich der Werteverteilungen.*

der Attribute bestimmt die Größe des Suchraums für die Heuristiken. Es betrifft andererseits auch die Komplexität der zu erwartenden Ergebnisse und damit potentielle Nutzbarkeit durch den Menschen.

Wie im Abschnitt 2.3.5.4 *Dimensionsreduktion* beschrieben, ist das Ziel die Suche nach einer möglichst kleinen Teilmenge von Attributen, die andererseits möglichst die relevanten Informationen über die Verteilung der Datensätze erhält. Unterschieden werden dabei zwei Verfahrensklassen: Verfahren zur *Merkmalsextraktion* operieren auf Datensätzen mit mehreren numerischen Attributen mit kommensurablen Skalen, d.h. solchen Skalen, zwischen denen arithmetische Operationen verwendet werden dürfen. Verfahren zur Merkmals*selektion* werden im allgemeinen auf Daten angewendet, die diese Voraussetzungen nicht erfüllen. Diese Verfahrensklasse beschreibt Heuristiken, in denen die Teilmengen der Attribute nach bestimmten Schemata getestet und verfeinert werden.

Da für das Klassifikationsproblem Attribute beliebigen Skalentyps zulässig sein sollen, lassen sich Verfahren zur Merkmalsextraktion auf *allen* Attributen nicht anwenden. Dementsprechend werden mit Smartstripes Verfahren für die Merkmalsselektion untersucht und gesteuert. Die Vereinheitlichung von Brown macht Smartstripes für sehr viele Verfahren für die Merkmalsselektion zugänglich.

4.2.2.2 Diskretisierung der Merkmale

Die Anforderung, dass das Verfahren Merkmale aus allen Skalentypen (nominal, ordinal und numerisch) verarbeiten muss, verlangt zusätzlich - wie schon für die erste vorgestellte Technik - eine Diskretisierung aller Merkmale in der Datentabelle in Partitionen. Dabei wird in Kauf genommen, dass bei numerischen (bzw. stetigen) Wertebereichen Information verloren geht. Dadurch wird jedoch erreicht, dass nominale Daten und ordinale Daten miteinander vergleichbar werden.
Zunächst unterscheiden wir zwischen den *Werten* eines Merkmals und den Datensätzen in der Datentabelle. Der *Wertebereich* eines Merkmals beschreibt die Menge aller gültigen Werte eine Tabellenspalten; die Datensätze sind repräsentiert durch die Tabellenzeilen. Falls die Werte sehr ungleich verteilt sind, kann es vorkommen, dass ein kleiner Wertebereich eines Merkmals eine große Zahl von Datensätzen umfaßt, oder umgekehrt. Wir definieren die Partitionen über dem Wertebereich eines Attributes so, dass jeder Wert genau eine Partition zugewiesen wird. Indirekt wird dadurch auch jedem Datensatz genau eine Partition zugeordnet. Wir benennen ein Attribut A_s und seinen Wertebereich synomym. Die Partition von A_s sei eine Folge von Teilmengen $A_{s1}, A_{s2}, \ldots$, so daß jede Teilmenge in A_s enthalten ist, daß A_s vollständig überdeckt wird, und kein Paar von Teilmengen gleiche Elemente enthält.
Für numerische Werte schränken wir ein, dass die Teilmengen stets Intervalle sind. Jede Teilmenge beginnt dementsprechend direkt nach ihrem Vorgänger. Für eine interaktive Änderung der Partition können die Intervallgrenzen verschoben werden, und es können Intervalle geteilt und zusammengefaßt werden; d.h. die Größe der Partition ist interaktiv veränderbar. Für nominale Werte existiert keine natürliche Anordnung. Der Anwender darf die Werte nach der Initialisierung daher beliebig umordnen, um beispielsweise Werte zusammenzufassen, die jeweils „gute“ Partitionen bilden könnten.
Wie schon in der KVMap Technik werden die Partitionen mit Histogrammen dargestellt. Damit der Anwender aber nicht beim Start des Programms sämtliche Partitionen selbst erstellen müssen, werden diese automatisch initialisiert. Um die mögliche Komplexität der Berechnung bei numerischen Werten zu reduzieren, wird zunächt ein equidistantes Binning nach der Methode von Shimazaki und Shinomoto [SS07]. Im allgemeinen ist dieses Binning noch nicht geeignet für eine statistische Auswertung der Merkmale. Die einzelnen Bins werden daher geclustert mit dem *k-Medoids*-Algorithmus. Im Gegensatz zur üblichen Anwendung eines Clusteringverfahrens operiert dieses nur auf der Werteverteilung jeweils eines Merkmals. Dieser Algorithmus wurde gewählt, weil er erstens sehr einfach ist, und zweitens tolerant gegenüber Ausreißern ist.
Die automatische Vorberechnung ist besonders dann notwendig, wo eine große Anzahl von Merkmalen für die Analyse vorbereitet werden muß. Allerdings kann man nicht erwarten, dass diese Initialisierung in allen Fällen optimal ist. Um flexibel zu bleiben, kann die Partition in der Feature Partition View interaktiv verändert werden. Verfeinerte Methoden für die Erzeugung guter Partitionen, wie etwa Brushing oder Clustering können in diesen Ansatz integriert werden.

Die Formel für die Transinformation 2.1 (siehe voriger Abschnitt) deutet an, wofür die Partitionierung genutzt wird. Im Prinzip testen alle automatischen Verfahren für die Merkmalsselektion die statistische Unabhängigkeit zwischen zwei Merkmalen. In der Statistik betrachtet man die Werte von Merkmalen als Ergebnisse von Zufallsprozessen, d.h. als „Er-

eignisse“. Zwei Zufallsereignisse E_1 und E_2 gelten dann als unabhängig voneinander, wenn die folgende Gleichung gilt für die Wahrscheinlichkeit p:

$$p(E_1) \cdot p(E_2) = p(E_1 \cap E_2) \tag{4.20}$$

Diese Beziehung ist unspezifischer - und damit allgemeiner - als jede Abhängigkeit, die durch ein analytisches Modell beschrieben kann. Eine statistische Abhängigkeit ist zunächst nur eine Aussage über Unterschiede in der Wahrscheinlichkeitsverteilung eines Ereignisses, abhängig von der Kenntnis über ein anderes Ereignis. Die statistische Abhängigkeit wird hier für die Identifizierung von Kandidatenattributen genutzt, gerade weil es so ein allgemeines Konzept ist. Mit ihr sind Aussagen über potentielle Beziehungen zwischen Attributen möglich, *bevor* ein analytisches Verfahren und damit eine Modellfamilie gewählt werden kann.

		A_s						
		1	2	3	...		n	Σ
A_r	1	f_{11}	f_{12}	f_{13}			f_{1n}	row_1
	2	f_{21}	f_{22}					row_2
	3	f_{31}		...				...
	...						...	...
	m	f_{m1}				...	f_{mn}	row_m
	Σ	col_1	col_2	...			col_n	N

Abbildung 4.14: *Die Kontingenztabelle ordnet die relativen Frequenzen der Verteilung der beiden Attribute A_r und A_s. Sind die beiden Attribute unabhängig, ist die relative Verteilung in allen Spalten und allen Zeilen gleich.*

Häufig verwendete Testverfahren für den Unabhängigkeitstest sind der *Chi-Quadrat-Test* und der äquivalente *G-Test*, die den Vorteil haben, dass es sich um verteilungsfreie, nichtparametrische Tests handelt. Die Testgröße des G-Test beruht auf der Transinformation, die ja für die Merkmalsselektion genutzt wird. Die Testgrößen eignen sich nur für Tests auf nominalen Merkmalen. Um numerische Merkmale gemeinsam mit nominalen Attributen zu testen, muss für die numerischen Merkmale eine Diskretisierung gefunden werden.
Hieraus ergibt sich eine der Schnittstellen für die Wechselwirkung zwischen den Qualitätsmaßen für die Selektion und der Diskretisierung der Merkmale. Mit der Diskretisierung der Merkmale ist es möglich, für jedes Paar von Attributen eine *Kontingenztabelle* zu berechnen (siehe Abbildung 4.14). Die Kontingenztabelle beschreibt die Häufigkeitsverteilung der Datenobjekte für alle Kombinationen von Wertkategorien der Merkmale. Die relativen Häufigkeiten $p(A_{1i}, A_{2j})$ dienen als Schätzer für die Wahrscheinlichkeiten. Wenn die Attribute A_1 und A_2 unabhängig sind, muß die folgende Gleichung für alle Kategorien i und j gelten:

$$p(A_{1i}, A_{2j}) = p(A_{1i}) \cdot p(A_{2j}) \tag{4.21}$$

In diesem Fall hat die Transinformation, anhand derer die Relevanz und Redundanz eines potentiellen Kandidatenattributes geschätzt werden, den Wert Null. Folgerichtig würde man ein statistisch unabhängiges Attribut eher nicht in die Auswahl für die folgende Analyse kommen.
Zu beachten ist jedoch, dass die Qualität des Tests stark von der Qualität der Diskretisierung abhängt. Bei einer schlechten Diskretisierung kann es geschehen, dass Wertemengen zusammengefaßt werden, deren Verteilung repräsentativ für den ganzen Datensatz ist. In diesem Fall könnte eine tatsächlich vorhandene Abhängigkeit zwischen den Attributen nicht erkannt werden. Würde man im Extremfall die Partitionen so erstellen, dass die Merkmalswerte *zufällig* den Teilmengen zuordnet werden, dann würde alle Unterschiede zwischen den einzelnen Werten innerhalb der Kategorien nivelliert werden.
Die vorgestellte Strategie koppelt daher die Verfahren für die Diskretisierung mit statistischen Testverfahren. Dass die wechselweise Beeinflussung dieser Verfahren jedoch über eine längere Analyse bessere Ergebnisse liefert, kann man aus drei Gründen annehmen:
Erstens ist das Ziel einer Partition - wie auch im allgemeineren Fall beim Clustering - Kategorien zu identifizieren, die in sich möglichst homogen sind, sich aber gleichzeitig möglich stark unterscheiden. Je besser eine Diskretisierung nach diesen Kriterien ist, desto wahrscheinlicher ist es, dass Abhängigkeiten zwischen zwei Attributen durch den statistischen Test auch erkannt werden.
Zweitens muß man ohnehin davon ausgehen, dass eine einzige Partitionen eines Attributes genügt, um alle Abhängigkeiten mit *allen* anderen Attributen zu finden. Es können Fälle auftreten, in denen mehrere Partitionen entwickelt werden müssen, um ein Attribut gegen alle anderen Attribute zu testen.
Drittens gilt, dass wenn zwei Merkmale stochastisch unabhängig sind, dann sind sie bezüglich der Verteilung aller ihrer Werte unabhängig. In diesem Fall existiert *keine* Diskretisierung, für die die Unabhängigkeitstests eine Abweichung berücksichtigen würden. Allgemeiner formuliert ist die gemessene Abhängigkeit niemals größer als die tatsächliche Abhängigkeit zwischen den Merkmalen.

4.2.2.3 Feature Partition View

Die *Feature Partition View* zeigt den Wertebereich jedes Merkmals in der Form eines horizontalen Histogramms (siehe Abbildung 4.13(1)). Die Balken des Histogramms sind dabei die Bins (bei numerischen Merkmalen) oder repräsentieren jeweils einen Datenwert (bei nominalen Daten). The Länge der Balken entspricht der relativen Größe des Bin. Bei nominalen Daten dürfen diese beliebig innerhalb des Histogramms verschoben werden.
Die Grundlinie der Histogramme (auf ihrer linken Seite) besteht aus einer Reihe von Tasten, die die einzelnen Teilmengen der Partition repräsentieren, die für eine Diskretisierung zusammengefaßt werden. Durch Ziehen der oberen oder unten Kante wird der Wertebereich entsprechend über die benachbarten Bins erweitert. Auf die gleiche Weise können Teilmengen zusammengefaßt werden. Zusätzlich dienen die Tasten als Filter. Einzelne Teilmengen können aus der aktuellen Analyse ausgeschlossen werden, um sich beispielsweise auf besonders ungewöhnliche Fälle zu fokussieren.

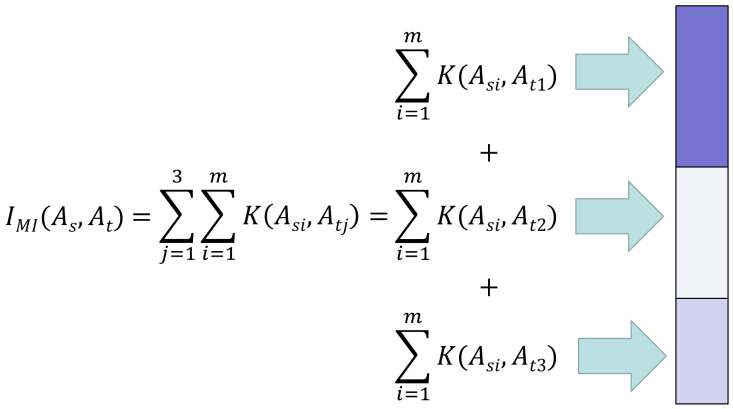


Abbildung 4.15: *Die Summenformeln für die Transinformation werden zerlegt. Die Summanden sind durch die einzelnen Zellen einer Spalte repräsentiert. Je größer die Summe, desto größer ist die lokale Abhängigkeit und desto dunkler der Farbton.*

4.2.2.4 Zerlegung der Summen

In diesem Abschnitt wird gezeigt, wie das Qualitätsmaß (siehe Gleichung 2.2) für ein bestimmtes Attribut A_s entlang der Partition zerlegt werden kann (siehe Abbildung 4.15). Die Qualitätsmaße beschreiben nur die Beziehung zwischen Attributen bezüglich *aller* Datensätze. Die wichtigste Komponente des Verfahrens ist daher die detaillierte Aufschlüsselung der Qualitätsmaße nach der Partition. Die Idee besteht darin, die Summenformeln so aufzuteilen, dass die Beiträge jeder Teilmenge der Daten einzeln darstellbar wird.
Jede Teilsumme repräsentiert eine Teilmenge der Datensätze, die durch die Partition definiert wird. Durch die Kombination der Gleichungen 2.1 und 2.2, kann man zeigen, welcher Teil des Qualitätsmaß entlang welches Attributes zerlegt werden kann: Der Relevanzterm Q_{Rel} besteht aus zwei verschachtelten Summen, deren Verschachtelung natürlich vertauscht werden kann.

$$Q_{Rel}(A_s) = \sum_{j=1}^{m} \underbrace{\sum_{i=1}^{n} K(A_{si}, A_{tj})}_{\text{Zerlegung nach } A_t} = \sum_{i=1}^{n} \underbrace{\sum_{j=1}^{m} K(A_{si}, A_{tj})}_{\text{Zerlegung nach } A_s} \tag{4.22}$$

Der Redundanzterm Q_{Red} besteht aus drei verschachtelten Summen, wobei nur der Merkmal A_s selbst in allen Termen auftaucht.

$$Q_{Red}(A_s) = -\beta \sum_{k=1}^{o}\sum_{j=1}^{n}\sum_{i=1}^{m} K(A_{si}, A_{kj}) = \sum_{i=1}^{m} \underbrace{-\beta \sum_{k=1}^{o}\sum_{j=1}^{n} K(A_{si}, A_{kj})}_{\text{Zerlegung nach } A_s} \tag{4.23}$$

Da man nicht annehmen kann, dass eine gemeinsame Partition für alle Merkmale existiert, darf nur das Merkmal A_s für die Zerlegung des Redundanzterms verwendet werden. Aus dem gleichen Grund kann die Summe für die bedingte Information Q_{Cond} nur nach den Partitionen von A_s oder A_r zerlegt werden.

Nun kann man definieren, welcher Teil des Qualitätsmaßes für welche Zerlegung genutzt werden kann. Der Anwender wird später zwischen zwei Zerlegungsstrategien wählen können:

1. Die Partition des Referenzmerkmals A_r wird für alle Merkmale verwendet. Der Vorteil dieser Zerlegung besteht darin, die Werte für alle Teilmengen der Partition über alle Merkmale A_s verglichen werden können. Ihr Nachteil besteht darin, dass sie nur für die Komponenten Q_{Rel} (siehe Abbildung 4.16a)) und Q_{Cond} (siehe Abbildung 4.16b)) genutzt werden. Vor allem sind sie nicht für die Gesamtqualität $Q(A_s)$ einsetzbar.

2. Das Qualitätsmaß jedes Merkmals A_s wird mit seiner jeweils eigenen Partition zerlegt (siehe Abbildung 4.16c)). Der Vorteil dieser Zerlegung besteht darin, dass man sie für die Gesamtqualität nutzen kann. Der Nachteil dieser Strategie besteht darin, dass die Teilmengen für unterschiedliche Merkmale nicht mehr verglichen werden können, da sie nichts miteinander zu tun haben müssen.

Durch die Zerlegung wird das Maß für die Gesamtqualität ergänzt durch eine Anzahl von detaillierteren Werten für jedes Merkmal A_s. Zwei Merkmale können durchaus eine identische Gesamtqualität haben; im Detail können sich die Beiträge einzelner Teilmengen jedoch unterscheiden. In machen Fällen kann eine einzige Teilsumme den Hauptteil für die Gesamtsumme enthalten. Gerade solche Abhängigkeiten, die nur einen Teil des Datensatzes ausmachen, können dennoch für die Klassifikation interessant sein. Lokale Muster, die eine genauere Ansicht wert sein könnten, werden der *Dependency View* sichtbar.

4.2.2.5 Dependency View

Die *Dependency View* (siehe Abbildung 4.13(2)) ist die zentrale Visualisierung in diesem Verfahren. Sie zeigt die Qualitätsmaße, die mit dem jeweils gewählten automatischen Verfahren für die Selektion der Merkmale berechnet werden. Es handelt sich um ein Matrix-basiertes Layout, in dem jede Spalte die Daten für eines der Merkmale darstellt. Da eine Spalte insgesamt alle Daten repräsentiert, entspricht eine Zelle jeweils einer Teilmenge.
Nach der ersten Zerlegungsstrategie (siehe voriger Abschnitt) ist die Anzahl und Größe jeder Zelle in einer Spalte definiert durch das Referenzmerkmal A_r und seine Partition (siehe Abbildung 4.16). Nach der zweiten Zerlegungsstrategie ist die Anzahl und Größe jeder Zelle definiert durch das Merkmal der jeweiligen Spalte. Die Höhe der Zelle ist proportional zur Größe der Teilmenge, die sie repräsentiert; sie stellt damit auch den relativen Einfluß auf die Gesamtqualität dar. Abhängig von der Zerlegungsstrategie kann der Anwender auch die einzelnen Komponenten Q_{Rel}, Q_{Red} or Q_{Cond} untersuchen.
Die Farbsättigung jeder Zelle wird durch den normierten Wert für jede Teilsumme der Zerlegung definiert. Niedrige Werte sind daher weiß und repräsentieren Teilmengen, die repräsentativ für alle Daten sein könnten. Das heißt, die Verteilung der Werte weicht kaum von der durchschnittlichen Verteilung ab. Hohe Werte sind tiefblau und repräsentieren Teilmengen, die stark vom Rest der Daten abweichen und daher auf eine lokale Abhängigkeit hindeuten kann.
Zusätzlich zeigt die Dependency View auch die Gesamtqualität für ein Merkmal als Bar-Chart über der jeweiligen Spalte. Die Gesamtqualität hängt dabei vom gewählten Verfahren

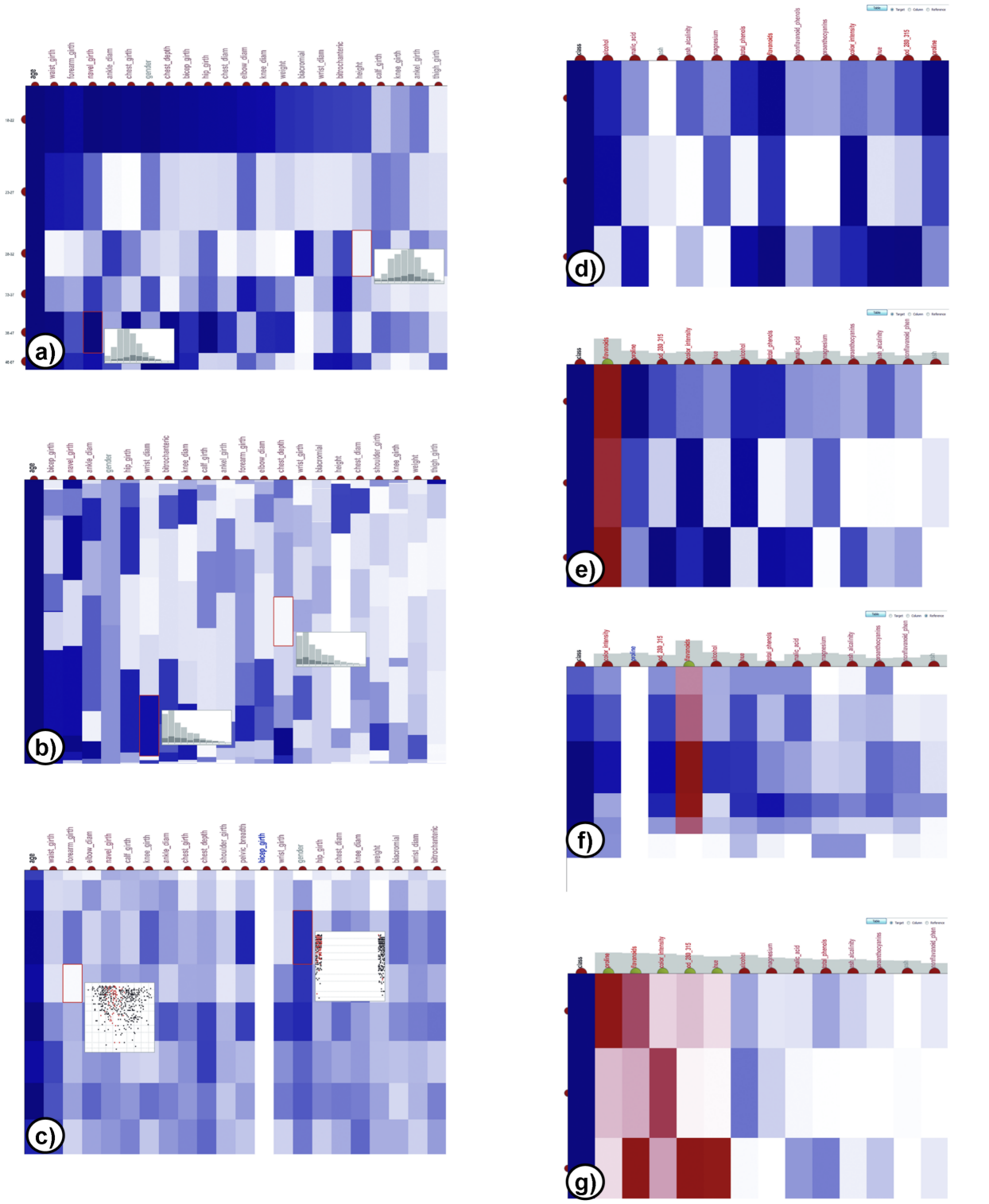


Abbildung 4.16: *Der Smartstripes Workflow. Jede Spalte kann unterteilt werden bezüglich der Partition des Targetattributes (a), bezüglich der eigenen Partition (b) und bezüglich der Partition eines gewählten Referenzattributes. (d) zeigt den ersten Blick auf die Daten. In Abbildung (e) wurden die Attribute nach der Relevanz sortiert und das wichtigste Attribut ausgewählt (rot unterlegt). (f) zeigt die Untersuchung von Relevanz und Redundanz bezüglich eines Referenzattribut. In (g) wurde schließlich eine Teilmenge der Attribute ausgewählt, die mit weiteren Verfahren im Detail untersucht wird.*

ab. Dabei ist die Höhe der Balken bezüglich des Maximum und Minimum normiert; das Resultat wird dadurch nicht verfälscht, weil es weniger auf die absolute Qualität als vielmehr auf den Qualitätsvergleich verschiedener Merkmale ankommt.
Das Referenzmerkmal A_r, mit dem die Qualität berechnet wird, wird mit dem abhängigen Attribut A_t initialisiert. Während das abhängige Attribut immer gleich bleibt - es definiert schließlich das Ziel der Klassifikation - kann das Referenzmerkmal während der Analyse gewechselt werden (in dem der Anwender auf die Spalte rechtsklickt). Dies ist sinnvoll, um Redundanzen zwischen potentiellen Kandidaten zu überprüfen, die nicht im Bezug das abhängige Merkmal A_t erkannt werden können. Im nächsten Abschnitt werden wir im Detail beschreiben, wie eine methodische Untersuchung dieser Abhängigkeiten unterstützt wird.
Die Dependency View enthält zusätzlich Tooltip-Visualisierungen für jede Zelle, die Details der Qualitätsberechnung der entsprechenden Teilmengen illustriert (siehe Abbildung 4.13(5)). Da jede Qualitätsberechnung im wesentlichen auf den Vergleich von Wertverteilungen hinausläuft, werden in den Tooltips entsprechende Verteilungen als Histogramme dargestellt. Dabei zeigt ein Histogramm eine Werteverteilung über alle Daten, ein zweites Histogramm die Werteverteilung über die Teilmenge. Nur für den Konditionalterm Q_{Cond}, der ja durch drei Merkmale berechnet wird, ist dieser Tooltip ein Streudiagramm, in dem die Teilmenge rot markiert wird (siehe Abbildung 4.16(c)).
Das Resultat der Merkmalsauswahl ist eine Teilmenge der Merkmale. Für diesen Prozess wird eine Kopplung zwischen Algorithmus und interaktiver Visualisierung etabliert. Einerseits kann der Anwender Merkmale selbst auswählen oder aus der Auswahl entfernen (durch Linksklick auf die jeweilige Spalte). Zusätzlich kann der Anwender auch Teilmengen aus der Analyse entfernen - etwa solche, für die eine Abhängigkeit schon bekannt und beschrieben ist. Nach jeder Veränderung der Auswahl oder der Filter wird eine automatische Neuberechnung der Qualitätsmaße ausgelöst. Wegen der Redundanzen zwischen den Merkmalen können sich auch alle Qualitätsmaße ändern. Um danach die besten Merkmale für die nächste Iteration zu finden, ist es unter Umständen notwendig, die Spalten neu zu sortieren. Durch die Sortierung kann der Anwender die Merkmale zusätzlich nach verschiedenen Metriken vergleichen. Das Sortieren ist der Teil der Technik, der die Visualisierung überhaupt erst skalierbar macht bezüglich der Anzahl der Merkmale in den Daten. Die Standardansicht auf die Matrix überdeckt den linken Teil eines virtuellen Bildschirmbereichs. Der virtuelle Bereich erstreckt sich so weit wie notwendig, um alle Merkmale unterzubringen; er wird aber nicht vollständig gezeigt. Bei der Sortierung werden die Merkmale der Qualität nach von links nach rechts absteigend geordnet. Die Standardansicht garantiert daher den Blick auf die besten Merkmale, der - falls notwendig - durch Scrolling immer noch erweitert werden kann.
Zusätzlich sind die Dependency View und Feature Partition View miteinander verbunden: Nach jeder Änderung an einer Partition wird die Qualität der Merkmale neu berechnet. Umgekehrt werden alle Histogramme in der Feature Partition View eingefärbt, wenn in der Dependency View eine Zelle selektiert wird, um anzuzeigen, wo die entsprechende Teilmenge in den Histogrammen liegt. Mit dieser Informationen können die Split-Punkte manuell verfeinert werden, die die Partitionen voneinander trennen.

4.2.2.6 Arbeiten mit Smartstripes

Eine Analyse mit SmartStripes beginnt mit der Definition eines abhängigen Merkmals A_t, das die Labels für die Klassifikation definiert. Dieses Merkmal ist das wichtigste Referenzmerkmal während der Merkmalsselektion und die meisten Berechnungen werden sich auf darauf beziehen. Im folgenden wird zunächst ein einfacher Workflow beschrieben, in dem das Referenzmerkmal nicht geändert wird. Anschließend beschreiben wir, wie durch das ändern des Referenzmerkmals systematische redundante Attribute untersucht werden können.
SmartStripes hat zwei grundsätzliche „Betriebsmodi". Im ersten Modus wird die Heuristik für die Selektion automatisch durchgerechnet; der Zweite ist ein „Mixed-Initiative"-Modus in dem zwischen automatischer und manueller Heuristik gewechselt werden kann. Im automatischen Modus verläuft die Merkmalsselektion wie „üblich": Alle Iterationen der Heuristik werden ohne Unterbrechung durchgeführt. Der Prozess endet, wenn eine maximale Anzahl der Merkmale erreicht, oder ein minimaler Qualitätszuwachs unterschritten wird. Die gewählten Merkmale werden genauso hervorgehoben, als wären sie manuell selektiert. Dieser Modus dient im wesentlichen der Inspektion und dem „Fine-tuning" des Endergebnis.
Im Mixed-Initiative-Modus berechnet die automatische Heuristik jeweils nur eine Iteration und hält danach an. Die Qualitätsmaße werden entsprechend aktualisiert und dem Anwender als Vorschlag angezeigt. Die entgültige Entscheidung trifft jedoch der Anwender selbst. Der kann den Vorschlag des Verfahrens, aber auch den Namen und die Bedeutung eines Merkmals Attributes und besonders die detaillierte Aufschlüsselung der Attribute berücksichtigen (siehe folgender Abschnitt). Die Auswahl eines Merkmals startet die nächste Iteration (siehe Abbildungen 4.16(d)+(e)).
Jedes Merkmal wird aus der Auswahlliste gelöscht oder hinzufügt; abhängig vom vorherigen Status. Auf diese Weise können auch die Heuristiken mit Forward- und Backward-Propagation einfach integriert werden. Überdies können die beiden Betriebsmodi im Wechsel genutzt werden: Es ist beispielsweise möglich ein automatisches Verfahren im Mixed-Initiative-Modus manuell zu initialisieren, und erst die letzten Iterationen automatisch zu berechnen.
Wegen des schrittweisen Vorgehens erlaubt der Mixed-Initiative-Modus auch Abweichungen von geradlinigen Workflow. Oft ist es interessant, Abhängigkeiten zwischen anderen Merkmalen als dem Targetattribut im Detail zu untersuchen (siehe Abbildung 4.16(f)). Im Vergleich zu einer Korrelationsmatrix, zeigt dieses Verfahren eine 1-zu-N-Beziehung zwischen dem Referenzattribut und allen anderen. Um beliebige Paare von Attributen in Bezug zu setzen, muß das Referenzattribut daher geändert werden. Das ist nützlich bei der Bewertung der von den automatischen Verfahren vorgeschlagenen Attribute. Beispielsweise wird das erste Attribut wegen seiner hohen Relevanz in die Auswahlliste hinzugefügt. Bevor dieser Vorschlag jedoch akzeptiert wird, kann der Anwender untersuchen, ob dieses wiederum von zwei oder mehr Attributen abhängt, die für die Analyse noch nützlichere Informationen enthalten könnten.
Weil die Abhängigkeiten zwischen allen Merkmalen einen vollständigen Graphen bilden, ist die Suche in den Graphen nicht linear. Abgesehen von einer Darstellung dieses Graphen, kann die Anordnung der Spalten der Dependency View die Organisation dieser Suche wirkungsvoll unterstützen: Erstens ist die Sortierung der Spalten zwar selbst automatisch, sie wird jedoch stets durch den Anwender ausgelöst. Auf diese Weise kann der Anwender be-

stimmte Merkmale im Blick behalten wenn nötig. Zweitens, wird die Sortierung für drei Gruppen von Attributen separat durchgeführt. Die erste „Gruppe“ ist das abhängige Attribut für die Klassifikation. Dieses wird am weitesten links dargestellt und niemals umsortiert. Diese Spalte dient damit als Ankerpunkt, um zum Hauptteil der Analyse zurückzugelangen. Rechts daneben sind die Attribute in der Auswahlliste, absteigend sortiert nach ihrer Qualität. Alle anderen, noch nicht auswählten Attribute befinden sich auf der rechten Seite des virtuellen Bildschirmbereichs, ebenfalls absteigend sortiert.
Zusammenfassend besteht der Workflow bei der Auswahl stets aus einem Hauptstrang, der verfolgt wird, wann immer das Referenzattribut auf das Attribut für die Klassifikation ist und einer Reihe von Nebenstrang, die für die detaillierte Untersuchung von Kandidaten für die Auswahl oder Entfernung genutzt werden. Üblicherweise wird der Hauptstrang wiederbetreten, sobald die Auswahlliste verändert wurde. Die Qualitätsmaße, die in der Visualisierung dargestellt werden, können jederzeit geändert werden.

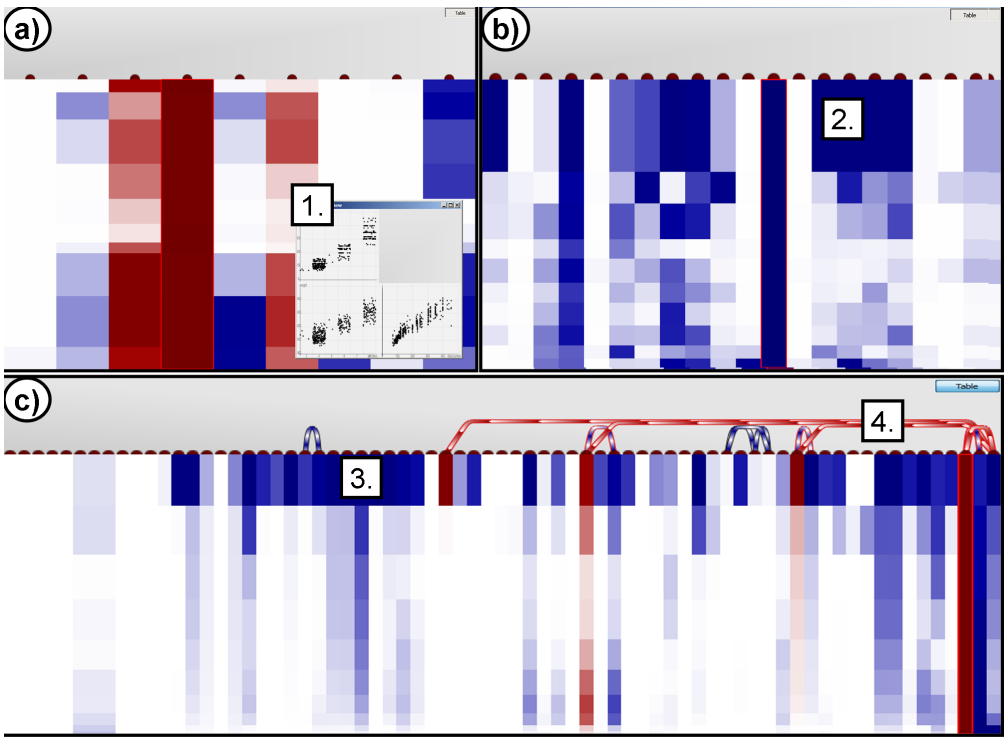


Abbildung 4.17: *Dieses Bild zeigt drei Visualisierungen mit den Ergebnissen des Unabhängigkeitstests. a) zeigt eine Darstellung des* Cars*-Datensatzes. Drei stark abhängige Attribute wurden für eine genauere Analyse (1.) mit der Scatterplotmatrix selektiert. b) zeigt eine Darstellung eines Datensatzes von Blutbildern. Als Referenzattribut der Cholesterolwert gewählt. Eine Reihe von Stoffen (Kupfer, Triglyceride) zeigt deneben ein ähnliches Profil der Abhängigkeit, was eine Analyse der Ähnlichkeit in dieser Gruppe rechtfertigt. Abbildung c) zeigt den US-Census Datensatz mit 70 Dimensionen. Auffällig ist hier der Einfluss der* No-Data *Werte auf viele Attribute (3.), der in den folgenden Iterationen eliminiert werden sollte. Darüber hinaus werden bestimmte Abhängigkeiten als Klammern oberhalb der Attribute gezeigt (4.). Diese Abhängigkeiten sind bereits vor der Analyse bekannt und sollten bei der Auswahl ebenfalls berücksichtigt werden.*

4.2.2.7 Lesen der Dependency View

Im Kern stellt das Verfahren Informationen gegenüber, die sich auf einzelne Teilmengen beziehen und solche, die sich auf Attribute beziehen. Die Dependency View stellt dar, wo die automatische Auswahl von Attributen sensibel auf eine unterschiedliche (Vor-)Auswahl von Teilmengen reagiert und umgekehrt. Wo auch immer die Qualitätsmaße entlang einer Spalte ungleich verteilt sind, wird die Filterung einer Teilmenge sehr wahrscheinlich die Ergebnisse der Attributauswahl beeinflussen.
Einige visuelle Artefakten stechen hervor und sind eine genauere Betrachtung wert. Die besten Kandidatenattribute sind üblicherweise jene, die sich als ununterbrochende blaue Linie darstellen. Alle Werte des Attributes tragen Information für die Klassifikation bei. Ebensogut kann es unterschiedlich starke Abhängigkeiten innerhalb des Wertebereichs eines Attributes geben. Werden die Qualitätsmaße entlang der Partition des Referenzattributs durchgeführt, bezieht sich jede Reihe von Zellen auf die gleiche Teilmenge der Daten. Bisweilen heben sich auch diese Reihen über mehrere Spalten hinweg deutlich vom Rest der Daten ab. Die entsprechenden Datensätze unterscheiden sich unter Umständen stark vom Rest der Daten. Abhängig von der Stärke der Abweichung, kann es sinnvoll sein, diese Datensätze getrennt zu untersuchen.
Teilmengen, die sich durch überdurchschnittlich helle oder dunkle Reihen von Zellen abheben können zudem genutzt werden, um die Partition des jeweiligen Referenzattributes zu verbessern. Eine gute Partition wäre möglichst klein, und würde dennoch das Gros der Information im Datensatz bewahren. Werden Teilmengen zusammenfasst, verhält sich die entstehende Teilmenge häufig ähnlicher zur Gesamtverteilung; das Maß für die Abhängigkeit wird entsprechend schwächer. Daher können Teilmengen, die eine starke Abhängigkeit zeigen, zusammengefasst werden, solange diese Abhängigkeit auch als dunkle Reihe sichtbar bleibt. Umgekehrt können Teilmengen, die durchgehend durch helle Zellen dargestellt werden eventuell unterteilt werden, um zu sehen ob die neue Partition mehr Details enthüllt.
Werden die Qualitätsmaße nach den Partitionen der Kandidatenattribute zerlegt, dann wird jede Spalte zu unterteilt wie ihr Attribut. Normalerweise können in dieser Ansicht keine Reihen identifiziert werden, weil jedes Attribut seine eigene Partition haben kann. Allerdings lassen sich in dieser Ansicht jene Kombinationen aus Teilmengen und Kandidaten finden, die für die Klassifikation lokal am besten geeignet wären. Anstatt nur jene Kandidaten zu bevorzugen, die eine „solide“ durchschnittliche Qualität zeigen, können diese abgewogen werden gegen Kandidaten, die eine starke Abhängigkeit auf einem kleinen Datenbereich zeigen. Weil praktisch alle Qualitätsmaße die „durchschnittlich guten“ Attribute in den Fokus rücken, fügten wir in zusätzliches Maß hinzu, nach dem die Spalten sortiert werden können. Es betrachtet nicht die Summe, sondern den maximalen Beitrag einer Teilmenge zum Qualitätsmaß eins Attributes, gewichtet nach der Größe der Teilmenge. Damit wird das Risiko minimiert, interessante, lokale Muster zu übersehen, die sonst aus dem sichtbaren Bereich entfernt werden könnten.
Vorsicht ist geboten bei der Interpretation, weil die Partitionen nicht notwendig optimal sind für eine Klassifikation: Muster, die sich in mehr als einem Attribut manifestieren, werden durch univariate Partitionen möglicherweise schlecht wiedergegeben. Dennoch ist man auch nicht darauf angewiesen, dass die Partitionen der Attribute optimal sind. Abhängigkeiten können auch dann noch sichtbar sein, wenn die Anzahl an Datensätzen hoch genug ist.

4.2.3 Evaluation

Für eine erste Evaluierung wurden echte Datensätze wie der US Microzensus verwendet. In diesem Datensatz sind mehrere Attribute redundant und die bereits dokumentierten Beziehungen wurden als „Ground-Truth“ genutzt. Starke, globale Abhängigkeiten können durch unser Verfahren, aber auch durch automatische Verfahren leicht erkannt werden. Aber auch eine Anzahl lokaler Abhängigkeiten, die Smartstripes sichtbar machen kann, konnten durch die Dokumentation des Datensatzes überprüft werden. Einige davon, wie die stark abweichenden Profile leerstehender Haushalte, erscheinen im Rückblick durchaus trivial. Merkmale mit einer gemischten Codierung - etwa ein numerisches Datum ergänzt um einen nominalen Wert wie „nicht verfügbar“ - hinterlassen ebenfalls eine gut erkennbare Spur in der Visualisierung. Andere Abhängigkeiten, wie die redundante Codierung des *Haushaltstyps* und der *Ausbildungsgrad des Haushaltvorstands* erfordern schon mehr Erfahrung mit den Daten, wenn man automatische Verfahren darauf vorbereiten muß. Die Exposition der Qualitätsmaße kann den Analysten von möglichen Fehlern bei der Analyse abhalten.

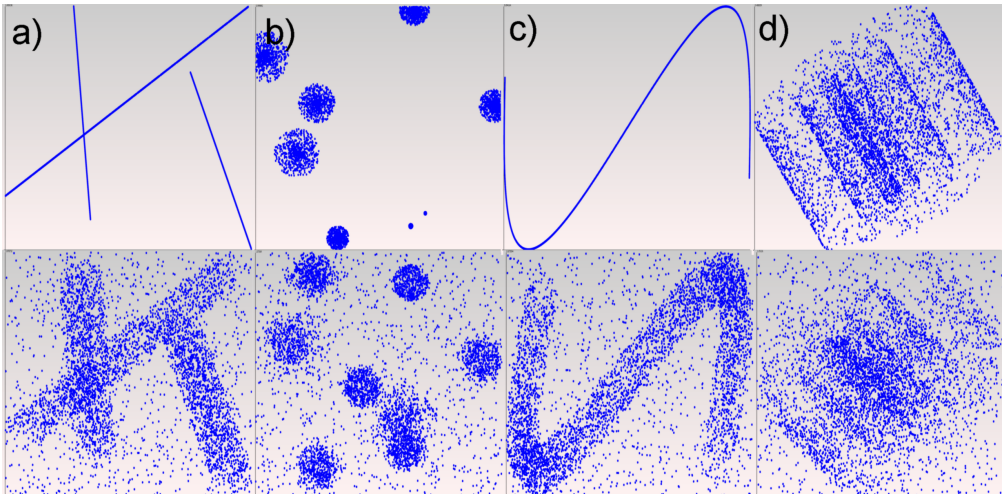


Abbildung 4.18: *Für den Test des Verfahrens wurden verschiedene Daten erzeugt, deren Abhängigkeit durch ein bestimmte Modell gegeben ist (oben). Zusätzlich wurden die Modelle mit einem bestimmten Anteil Rauschen versehen (unten, 20 Prozent Rauschen). Von jeder Modellklasse können zufällig verschiedene Varianten erzeugt werden. Die Position der Linien in a) und die Position und Größe der Cluster in b) kann variiert werden. Die Sinuskurve (c) und die sogenannte „Swiss Roll“ (d) können im Raum beliebig gedreht werden. Die „Swiss Roll“ ist dabei ein Beispiel für einen dreidimensionalen Datensatz.*

4.2.3.1 Blindstudie

Darüber hinaus wurde geprüft, ob mit Smartstripes auch noch stark verrauschte bzw. multivariate Abhängigkeiten in den Daten erkannt wurden. Multivariate Abhängigkeiten sind deshalb problematisch, weil in den Qualitätsmaßen maximal drei verschiedene Attribute Q_{Cond} berücksichtigt werden. Bestimmte Muster könnten daher übersehen werden.
Der Test war eine Blindstudie mit künstlichen Datensätzen die mit einem kontrollierten Anteil Rauschen versehen wurden. Für jeden Test wird die Größe des Samples festgelegt, d.h.

die Anzahl der Datensätze in der Tabelle. Schließlich werden mehrere Modelle definiert, die jeweils eine Abhängigkeit zwischen zwei oder mehr Attributen beschreiben. Die Datensätze der Tabelle werden den Modellen gemäß zufällig erzeugt. Ein beliebiges Paar von Attributen, die nicht nach dem gleichen Modell definiert wurden, ist unabhängig voneinander. Bei der Erzeugung kann auf zwei Arten Rauschen hinzugefügt. Erstens kann ein Prozentsatz der Datensätze angegeben, die nicht nach den Modellen definiert sondern rein zufällig erzeugt werden. Zusätzlich können auch Datenpunkte der Modelle - bei numerischen Attributen - um einen kontrollierten Betrag verzerret werden.
Für den Test werden die Attribute zufällig umsortiert. Der Test selbst besteht in der Aufgabe, die Attribute so gruppieren, dass genau jene Attribute, die nach dem gleichen Modell beschrieben sind, der gleichen Gruppe zugeordnet werde. Die Frage ist daher, für welche Modelle und für welchen Grad Rauschen, diese Gruppierung mit Hilfe des Verfahrens noch vorgenommen werden kann.
Die Performance der nicht verrauschten Daten war in allen Tests unabhängig von der getesten Modellklasse (siehe Abbildung 4.18). Dies ist dadurch zu erklären, dass die meisten Varianten dieser Modelle sich deutlich von unabhängigen Daten abheben. Dies trifft übrigens auch auf Modelle zu, die Abhängigkeiten in mehr als zwei Attributen beschreiben und in einem Streudiagram gut zu erkennen sind. Je mehr Rauschen die Daten enthalten, desto schwieriger ist die Unterscheidbarkeit abhängiger und unabhängiger Attribute (siehe Abbildung 4.19). Rauschen kann jedoch gut kompensiert werden durch die höhere Anzahl der Samples, denn diese Erhöhen die Sensitivität der statistischen Verteilungstests: Hohe Abweichungen von der erwarteten Verteilung bei vielen Samples fallen stärker ins Gewicht als eine Vergleichbar hohe Abweichung bei einer kleinen Anzahl von Samples. Solange genügend Daten zur Verfügung stehen, wäre damit jeder beliebig hohe Anteil (< 100 %) Rauschen zu kompensieren.
Ein Grund für die guten Ergebnisse kann aber auch darin bestehen, dass einfaches Rauschen kein guter Distraktor ist, um künstliche Abhängkeiten zu maskieren. Echte Daten wurden auf eine ähnliche Weise modifiziert, um zu überprüfen, ob sie einen besseren Distraktor darstellen, was sich auch bestätigte. Die Aufmerksamkeit der Anwender ist typischerweise gerade durch die stärksten - oftmals nur bivariaten - Abhängigkeiten gelenkt. In einigen Fällen konnten die dominanten Abhängigkeiten mit einer bestimmten Teilmenge assoziiert werden. Durch Entfernen dieser Teilmenge aus der Analyse konnten schwächere Beziehungen sichtbar gemacht werden. Die Anwender bemerkten jedoch auch, dass Teilmengen nicht immer präzise zu entfernen waren, weil die Partitionen einzelner Attribute dafür zu unflexibel seien. Eine freiere Definition dieser Abhängigkeiten wäre daher eine mögliche Verbesserung dieses Verfahrens.

4.2.3.2 Beschränkungen bei der Anwendung von SmartStripes

Wie bei allen Techniken, die auch statistischen Testgrößen basieren, besteht auch hier die wichtige Einschränkung, dass die Ergebnisse der Tests instabil werden, wenn die Anzahl der Samples zu klein ist. SmartStripes macht hier keine Ausnahme. Viele frei verfügbaren Daten enthalten weniger als tausend Datensätze. Valide Resultate sind damit nicht zu erzielen. Nach der Zerlegung der Testgröße entlang von zwei oder drei Attributen, kann es sein, dass ein einzelner Summand sich nur auf ein paar Samples bezieht. Das ist mithin auch der

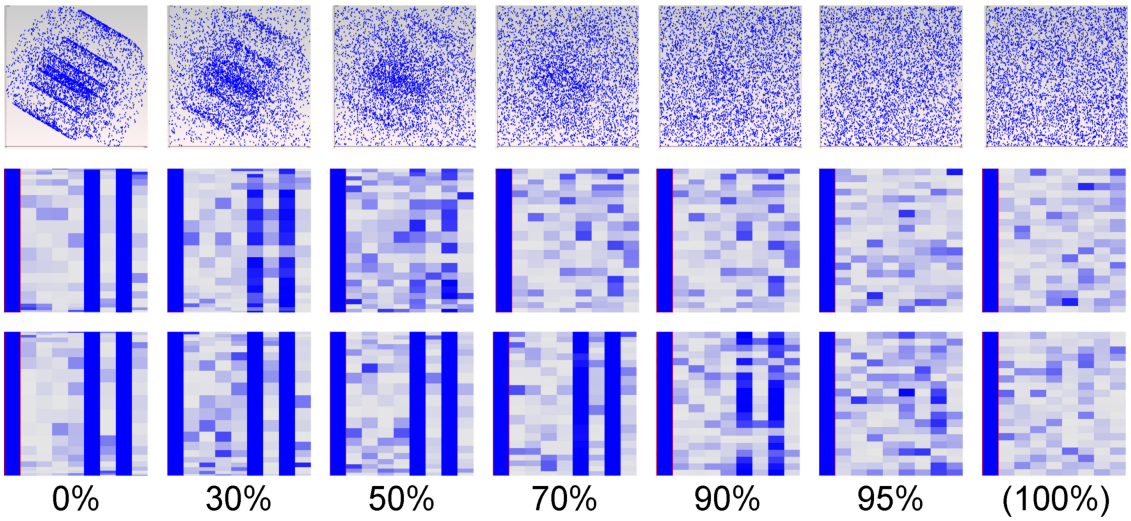


Abbildung 4.19: *Hier wird der Unabhängigkeitstest eine künstliche Datentabelle mit der „Swiss Roll“ in unterschiedlich verrauschten Varianten überprüft. Die unteren beiden Reihen zeigen dabei die Visualisierung der Testergebnisse mit 5000 Datenobjekten (mitte) und mit 100000 Datenobjekten (unten). Nur drei der neun Spalten der Datentabelle sind vom gewählten Referenzattribut abhängig. Generell hat das Rauschen einen negativen Einfluß auf die Unterscheidbarkeit der abhängigen und unabhängigen Spalten des Datensatzes. Positiv ist zu vermerken, dass dieser Einfluß durch eine hohe Anzahl von Samples kompensiert werden kann: Bei 100000 Samples ist können durch den Unabhängigkeitstest selbst bei 90 Prozent Rauschen noch die abhängigen Spalten identifiziert werden, was selbst durch den Scatterplot nicht mehr möglich wäre.*

Grund, warum automatische Verfahren für die Attributselektion selten mit multivariaten Abhängigkeitstests verwendet werden.
Weil das Verfahren interaktiv bleiben soll, existiert auch eine obere Schranke für die maximale Anzahl an Samples. Glücklicherweise kann ein Großteil der Berechnungen vorberechnet werden. Die aufwändigsten Berechnungen sind die Änderung des Targetattributes und die Änderung von Partitionen. Selbst ohne ein DBMS können Tabellen mit vier Millionen Einträgen (Samples x Attribute) noch interaktiv untersucht werden.
Die Technik ist durch den Mixed-Initiative-Modus sehr flexibel, was die Möglichkeiten für die Analyse erhöht, aber auch entsprechend komplex zu bedienen sein kann. Partitionen und Referenzattribut können jederzeit geändert werden, wobei der Anwender selten alle diese Informationen gleichzeitig berücksichtigt. Die automatische Optimierung der Auswahlliste ist nur eine von mehreren Möglichkeiten, bestimmte Teilprozesse wieder zu automatisieren. Das gleiche gilt etwa auch für die Optimierung der Partitionen, was die Freiheitsgrade für eine einfachere Bedienung reduzieren würde.

4.2.4 Kopplung und Rückkopplung von Attributauswahl und Data-Mining Verfahren

Am Anfang dieses Abschnitts wurde das vorgestellte Verfahren als Lösung für zwei Aufgaben motiviert. Die Selektion von Attributen für die folgenden Schritte der Analyse durch Data-Mining Verfahren oder Visualisierungen, und die Suche nach „guten“ Partitionen für die Attribute. Je besser die Partitionen die Ähnlichkeiten und Unterschiede in den Daten wiedergeben, desto trennschärfer werden auch die Unabhängigkeitstests für die Attributselektion. An mehreren Stellen wurde jedoch schon darauf hingewiesen, dass auch die Suche nach „guten“ Partitionen nicht nur entlang univariater Verteilungen erfolgen müßte, die relativ wenige Gestaltungsmöglichkeiten bieten. Mit dem gleichen Ziel operieren Clusteringverfahren auf multivariaten Verteilungen, wodurch natürlich auch multivariate Muster untersucht werden können.
Jedoch muß auch für Clusteringverfahren eine „gute“ Auswahl von Attributen gefunden werden. Verkürzt scheinen sich die beiden Aufgaben daher eigentlich als „Henne-Ei“-Problem darzustellen. Dies ist jedoch im allgemeinen deshalb nicht der Fall, weil keineswegs nur ein optimales Clustering die Unabhängigkeitstests verfeinern könnte und weil ebensowenig nur eine optimale Attributselektion die Vorausetzung für ein brauchbares Clustering liefert. Die Verfahren lassen sich insbesondere auch dadurch verbessern, dass möglichst viele bereits bekannte Ergebnisse für den jeweils folgenden Schritt genutzt werden können. Im folgenden wird gezeigt, wie diese Ergebnisse als synthetische Attribute repräsentiert werden und in Smartstripes weiterverwendet werden können.

4.2.4.1 Synthetische Attribute

Mit synthetischen Attributen kann man die Einschränkung umgehen, dass mit Smartstripes stets nur die Abhängigkeiten zwischen zwei bis maximal drei Attributen dargestellt werden. Synthetische Attribute repräsentieren Zwischenergebnisse, die mit verschiedenen Analysetechniken - Clusteringverfahren, Verfahren zur Dimensionsreduktion, aber auch einfache Datentransformationen - in vorangangenen Schritten der Analyse erzielt wurden. Dadurch können synthetische Attribute insbesondere die Informationen von jenen Attributen zusammenfassen, aus denen sie berechnet wurden.
Es spricht nichts dagegen, diese neuen Attribute in den folgenden Schritten ebenso zu behandeln, als wären sie Attribute des Originaldatensatzes. Für die Visualisierung mit Smartstripes handelt es sich dabei um eine Erweiterung um zusätzliche Spalten. Besonders Clusteringergebnisse bieten qua Definition eine gute natürliche Partition der Daten, auf deren Basis weitere Abhängigkeiten sensitiv untersucht werden können. Dass dabei Details einzelner Datensätze verloren gehen, wird dabei zugunsten einer höheren Informationsdichte in Kauf genommen.
Wenn nach *einer* Iteration mehrere Attribute auf ein synthetisches Attribut reduziert werden, dann ist es im Prinzip auch möglich, in mehreren Schritten rekursiv beliebig viele Attribute einzubeziehen und deren Abhängigkeiten zu untersuchen. Dies ist der zentrale Ansatzpunkt, um der Komplexität der Analyse hochdimensionaler Daten zu begegnen. Für jede Iteration können dabei auch Verfahren gewählt werden, die nur jeweils wenige Attribute

untersuchen können, das gilt im Besonderen für den Einsatz von Visualisierungstechniken. Abbildung 4.20 deutet an, dass diese Verdichtung eine Hierarchie von Attributen erzeugt. Da synthetische Attribute mehrdimensionale Abhängigkeiten repräsentieren, lassen sich so indirekt auch entsprechend komplexere Zusammenhänge erforschen, welche in den ursprünglichen Attributen u.U. nicht sichtbar wären.
Eine Staffelung dieses Prozesses wurde beispielsweise von Bishop et al. [BT98] oder auch Yang et al. [YWRH03] vorgeschlagen. Beide Ansätze operieren auf numerischen Daten. Bishop et al. konstruieren dafür semi-automatisch eine Hierarchie von Clustern, wobei für jede Stufe der Hierarchie ein neues Modell konstruiert wird. Yang et al. konstruieren eine Hierarchie von Attributen durch Clustering, die manuell verfeinert werden kann. Diese Hierarchie wird in diesem Ansatz dargestellt um geeignete Attribute für die Visualisierung auszuwählen.
Der hier vorgestellte Ansatz operiert im Unterschied dazu nicht allein auf numerischen, sondern auch auf nominalen Daten, da Abhängigkeiten in Smartstripes über die Entropiemaße definiert werden. Dies erlaubt es, beliebige Attribute miteinander in Bezug zu setzen.

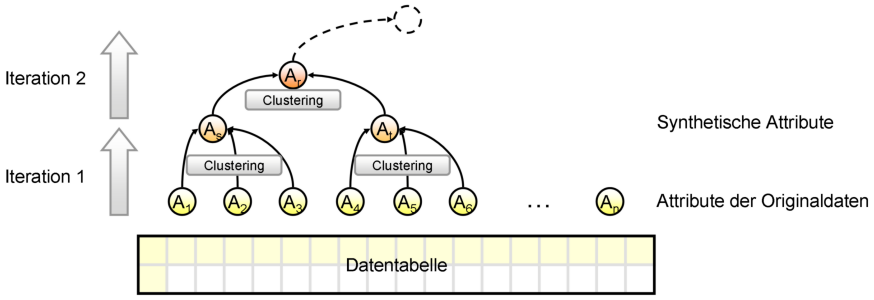


Abbildung 4.20: *Über mehrere Iterationen der Analyse findet eine Informationsverdichtung statt. Die synthetischen Attribute konsolidieren dabei die Zwischenergebnisse, die durch verschiedene Analyseverfahren und Modelle erstellt werden können. Damit wird das Problem der Skalierbarkeit einzelner analytischer Verfahren teilweise entschärft. Clusteringverfahren fassen dabei die Informationen mehrerer Attribute zu einem neuen, diskreten Attribut zusammen, das in folgenden Iterationen weiter kombiniert werden kann.*

Diese hierarchische Staffelung erlaubt es zusätzlich auch, beliebige Verfahren - und damit auch beliebige Modelle - miteinander zu kombinieren. Beispielsweise muß ein einzelnes Clusteringverfahren nicht in jedem Unterraum der Daten die besten Ergebnisse erzeugen. Innerhalb der Iterationen können beliebige Verfahren gewählt werden, wobei die synthetischen Attribute die „lingua franca“ bei der Kombination neuer Ergebnisse darstellen. Die neuen Attribute vergrößern, unabhängig davon wie sie erzeugt wurden, auch das Repertoire der Attribute für die Konstruktion von prädiktiven Modellen. Stehen nur die ursprüngliche Attribute $A_1, \ldots, A_{dim}$, dann muß das Klassifikationsmodell *classifier* direkt auf einer Teilmenge dieser Attribute operieren. Betrachtet man synthetische Attribute als Modelle Ψ aus anderen Verfahren, die sich direkt oder indirekt aus den ursprünglichen Attributen ableiten, dann hätte das kombinierte Klassifikationsmodell beispielsweise folgende Form:

$$classifier\left(A_1, \Psi\left(A_2, A_3\right)\right) \tag{4.24}$$

4.2.4.2 Iterative Verfeinerung

Durch die Entzerrung der Modellkonstruktion auf mehrere Iterationen und mehrere Verfahren können hybride Modelle erstellt werden, die sich besser an unterschiedliche Muster in den Daten anpassen lassen. Da komplexe Modelle schrittweise aus einfacheren Elementen konstruiert werden, bleibt der Prozess selbst überschaubar einfach. Smartstripes dient als Ankerpunkt für die Bewertung neuer Ergebnisse, d.h. synthetischer Attribute, für die grobe Untersuchung von Abhängigkeiten und für die Auswahl der Attribute für den nächsten Schritt der Analyse.
Bis auf die genannten Vorraussetzungen gibt es wenig Einschränkungen darüber, welche Verfahren innerhalb dieser Strategie für die Konstruktion von Modellen und synthetischen Attributen genutzt werden. Insbesondere können dabei sowohl automatische, semi-automatische aber auch visuell-interaktive Verfahren in diesen Prozess integriert werden. Die Verfahren, durch die neue synthetische Attribute erzeugt werden, können dabei so einfach sein wie Brushing, das bereits im Konzept genutzt wurde, um ein Klassifikationslabel zu erzeugen, dass mit den Originaldaten verglichen werden konnte. Mittelbar können so die Ergebnisse sehr unterschiedlicher Verfahren miteinander untersucht werden.

4.2.5 Vorteile von Smartstripes

Die Stärken der vorgestellten Strategie lassen sich folgendermaßen zusammenfassen:

- **Einfaches, allgemeines Modell für die Abhängigkeit zwischen Attributen:** Durch den Rückgriff auf parameterfreie, statistische Tests wird Abhängigkeit und Unabhängigkeit sehr allgemein, und unabhängig von vorgegebenen analytischen Modellen untersucht. Als Stand-alone Verfahren wäre dies eine Schwäche. Bei die Kopplung des Testverfahren mit anderen analytischen Techniken für die detaillierte Untersuchung der Abhängigkeit, ermöglicht der statistische Test jedoch eine Untersuchung *vor* der Wahl eines spezifischen Modells.

- **Skalierbarkeit bezüglich der Anzahl der Attribute des Datensatzes:** Durch das Matrix-Layout genügt eine Spalte mit wenigen Pixeln Breite, um die relevanten Informationen darzustellen. Dies ermöglicht die Darstellung mehrerer hundert Attribute auf einem Standard-Desktop.

- **Skalierbarkeit bezüglich der Anzahl der Datenobjekte:** Die Berechnungskomplexität steigt proportional mit der Anzahl der Datenobjekte. Durch die Darstellung aggregierter Daten - wie auch in der KVMap - bleibt die visuelle Komplexität unabhängig von dieser Größe. Der wichtigste Vorteil ist, dass die Sensitivität des statistischen Testverfahrens mit der Anzahl der Datenobjekte *zunimmt*. Je höher die Anzahl der Datenobjekte, desto sicherer werden Abweichungen von der Unabhängigkeitshypothese erkannt. Dies ist besonders in den Fällen relevant, wenn eine Abhängigkeit im hochdimensionalen Raum nach der Projektion auf nur zwei Achsen nur noch schwach zu erkennen ist.

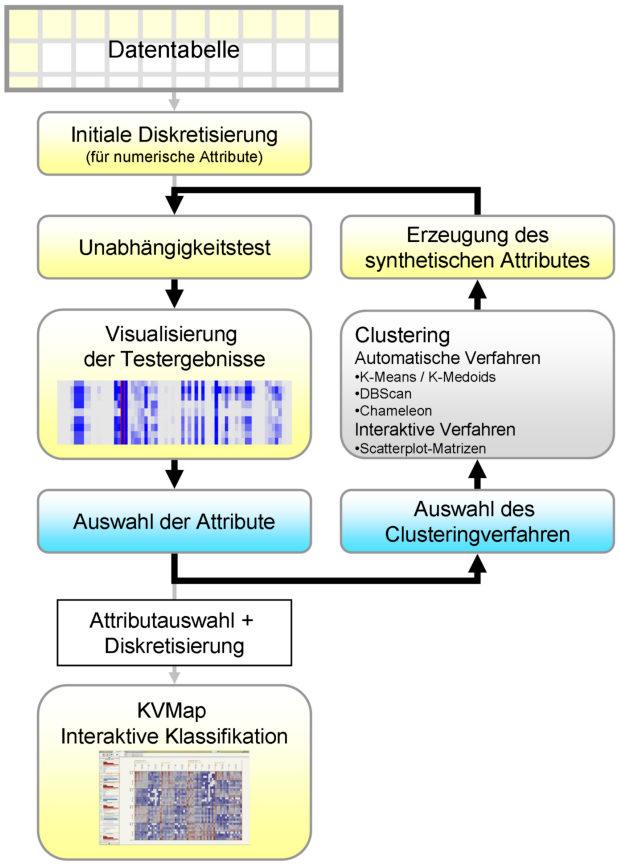


Abbildung 4.21: *Die hier vorgestellte Strategie für die Suche nach guten Attributen und Diskretisierungen ist ein iterativer Verfeinerungsprozeß. Die statistischen Unabhängigstests werden genutzt, um eine Kandidatenmenge abhängiger oder unabhängiger Attribute einzugrenzen, bevor ein Modell gewählt werden kann. Die Clusteringverfahren dienen zur Zusammenfassung von Informationen aus mehreren Attributen. Die Information kann so über mehrere Iterationen stufenweise verdichtet werden.*

- **Einheitliche Konsolidierung von Ergebnissen unterschiedlicher Verfahren:** Durch die Repräsentierung der Ergebnisse von Verfahren als synthetische Attribute der Datensätze wird die Strategie sehr flexibel. Zudem bietet diese Repräsentierung die Möglichkeit alle Teilergebnisse in Bezug zu setzen, da Abhängigkeiten zwischen ihnen direkt untersucht werden können.

4.3 Ergebnisse & Diskussion

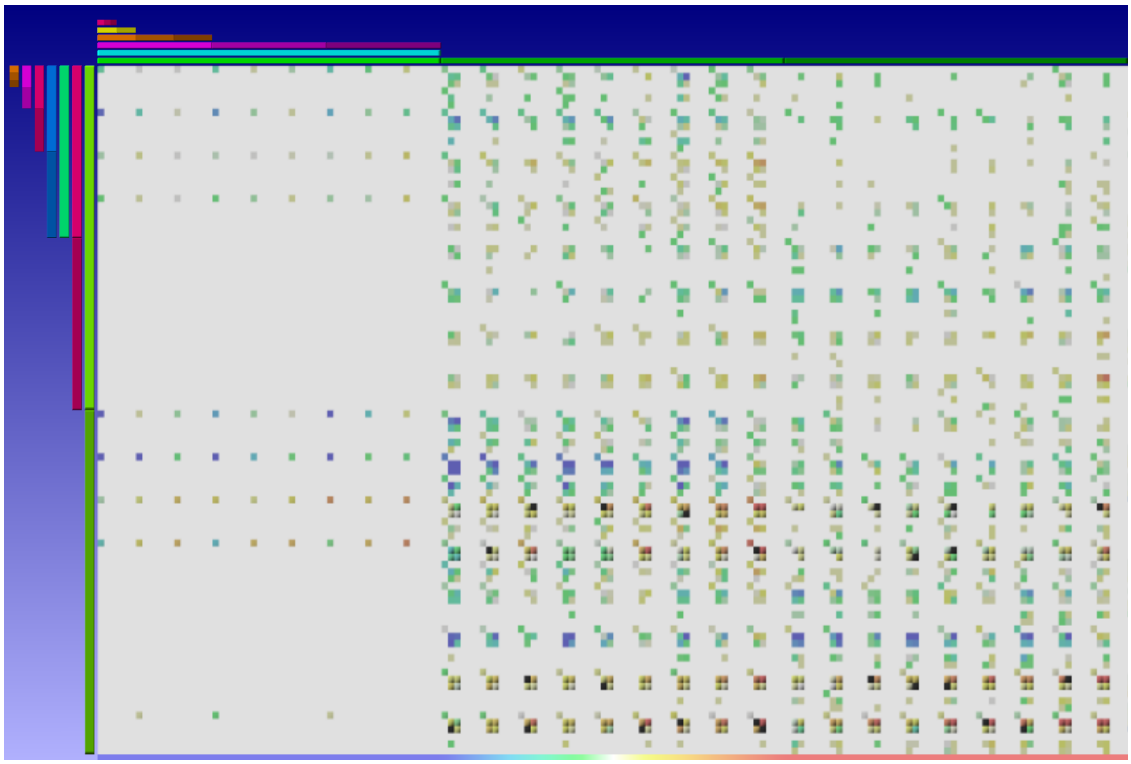


Abbildung 4.22: *Diese KVMap stellt elf Attribute des US-Zensus für 20.000 Haushalte dar, wobei jedes dieser Attribute in zwei oder drei Partitionen unterteilt wird. Das abhängige Attribut, mit dem die Profile korreliert werden, beschreibt, ob ein Haushalt Kinder hat oder nicht. Auch wenn andere Visualisierungstechniken mehr Dimensionen darstellen können (etwa* Parallelkoordinaten*) handelt es sich, soweit bekannt, um die einzige Visualisierung, die Bezüge zwischen zehn und mehr Dimensionen auf präattentiv wahrnehmbare Merkmale des Bildes abbilden kann. Dies liegt daran, dass auch die Überlagerung verschiedener Frequenzen zu einem Sinneseindruck integriert werden kann. Die Zusammenhänge sind jedoch zu komplex um direkt lesbar zu sein.*

Die hier vorgestellte Visualisierungstechnik KVMap wurde nach den Anforderungen entworfen, dass möglichst viele Dimensionen gleichzeitig darstellbar sind, besonders aber, dass die Bezüge, die als Muster wahrgenommen werden können, möglichst viele Attribute umfassen können. Andere Visualisierungstechniken können mehr Dimensionen direkt darstellen, wie beispielsweise Parallelkoordinaten [Sii00, JFLC08] oder auch tabellenbasierte Visualisierungen, wie die *Table-Lens* [RC94]. Allerdings ist es nicht möglich, auch hochdimensionale Bezüge zwischen den einzelnen Dimensionen direkt wahrzunehmen - sieht man von Ausnahmefällen ab, in denen etwa große Ähnlichkeit zwischen den Werten mehrerer Dimensionen sichtbar wird.
Abhängig vom Anzeigegerät, dem menschlichen Sichtfeld und der Partitionierung der einzelnen Attribute lassen sich zwischen zehn und zwanzig unabhängige Dimensionen gleichzeitig darstellen. Dies sind zwar weniger Dimensionen, jedoch ist es mit der *KVMap* möglich, auch entsprechend hochdimensionale Zusammenhänge als Muster sichtbar zu machen. Ein Grundproblem der visuellen explorativen Datenanalyse besteht darin, dass Zusammenhänge, die

komplexer sind als die Attribute, die gleichzeitig in Beziehung gesetzt werden können, unter Umständen schwer oder gar nicht sichtbar sind, wenn nur wenige Attribute von vornherein ausgeschlossen werden können.
Die Sichtbarkeit der Relevanz und auch der Irrelevanz eines Attributs ist potentiell abhängig von der Wahl der Attribute, die gleichzeitig analysiert werden. Das bedeutet, dass

- eigentlich relevante Attribute nicht identifiziert werden könnten, weil die Attribute, mit denen sie gemeinsam einen Zusammenhang beschreiben, nicht ausgewählt wurden.
- eigentlich irrelevante Attribute nur deshalb die Visualisierung dominieren, weil sie nur im Kontext mit noch weniger relevanten Attributen verglichen werden.

Die Möglichkeit, zehn- und höherdimensionale Zusammenhänge sichtbar zu machen entschärft dieses Problem [1]. Beispielsweise ist es möglich, aus einer Visualisierung mit zehn Dimensionen, effizient die relevantesten Attribute für die Charakterisierung des Targetset zu identifizieren. Bei der Verwendung der KVMap im Rahmen einer Analyse - ohne Voraussetzung von Vorwissen - hat sich folgende Strategie als sinnvoll erwiesen:

1. Auswahl weniger (vier bis sechs) Attribute für die Visualisierung
2. Interaktive Partitionierung dieser Attribute mit möglichst vielen Partitionen. Dies ist deshalb notwendig, weil optimale Partitionen a-priori nicht bekannt sein müssen, und durch diese Strategie möglichst wenig Information über die Originaldaten erhalten bleibt. Alternativ kann auch eine automatische Partitionierung gewählt werden.
3. Sukzessive Verkleinerung der Partitionierung. Wo sich Muster bezüglich verschiedener Partitionen wiederholen, können diese Partitionen zusammengefasst werden, ohne dass viel Informationen verloren geht. Durch die Verkleinerung einer Partitionierung, reduziert sich die Anzahl der Zellen in der Visualisierung, wodurch man Platz für weitere Attribute oder eine Verfeinerung der Partitionierung an anderer Stelle gewinnt.
4. Hinzufügen jeweils eines neuen Attributs.

Nach verschiedenen Kriterien können Attribute aus der Visualisierung entfernt werden, um Platz zu schaffen für noch nicht untersuchte Attribute. Ein Kriterium besteht zum Beispiel darin, iterativ jeweils das Attribut zu entfernen, das am wenigsten relevant ist. Mit dem Smartstripes-Verfahren und bei der Konstruktion der Entscheidungsbäume wird diese Relevanz quantitativ bestimmt. Sie kann innerhalb der Visualisierung auch sichtbar gemacht werden: Ein Attribut ist dann wenig relevant, wenn sich die Muster in der Visualisierung durch Hinzufügen oder Entfernen des Attributes kaum verändern. Nacheinander können im Prinzip alle Attribute einer Datentabelle gestestet werden, um eine Teilmenge der Attribute zu finden, mit der ein guter Prädiktor gefunden werden kann.

[1] Dennoch sei hier angemerkt, dass dieses Problem nicht prinzipiell gelöst wird. Dies wäre nur möglich durch eine erschöpfende Untersuchung aller Kombinationen von Attributen. Dies ist im Allgemeinen auch durch automatische Verfahren nicht möglich.

Die Möglichkeit, komplexe Zusammenhänge darzustellen, macht die visuelle Bewertung der Rolle eines einzelnen Attributes sicherer, verursacht aber letztlich das Problem, das das Konzept dieser Arbeit motiviert: Je komplexer die Zusammenhänge, desto mehr Informationen muss der Anwender für deren Beschreibung suchen und behalten. Im Falle der *KVMap* sind dies die Namen der Attribute, die Beschreibung der relevanten Partitionen und deren Abhängigkeiten. Die vorgestellten Verfahren für die Beschreibung dieser Muster sind dabei unterschiedlich mächtig. Während die Minterme nur Disjunktionen von charakteristischen Partitionen der relevanten Attribute darstellen, können im beschriebenen Entscheidungsbaum komplexere Ausdrücke konstruiert werden.
Bei der Anwendung des iterativen Feedback kommen in der Praxis alle Fälle vor, die in Abschnitt 3.2.3.1 beschrieben werden. Durch den visuellen Abgleich zwischen Referenzdaten und Feedback läßt sich nicht nur die Qualität des Modells selbst, sondern auch die Qualität des Verfahrens einschätzen. In der Praxis ergibt sich beim Feedback zwischen Mensch und Maschine entweder eine positive Rückkopplung, in der Anwender die durch das Modell vorgeschlagenen Ergänzungen des selektierten Musters annimmt oder eine negative Rückkopplung, in deren Folge der Anwender das verwendete Verfahren prüfen, wenn nicht sogar austauschen muss.

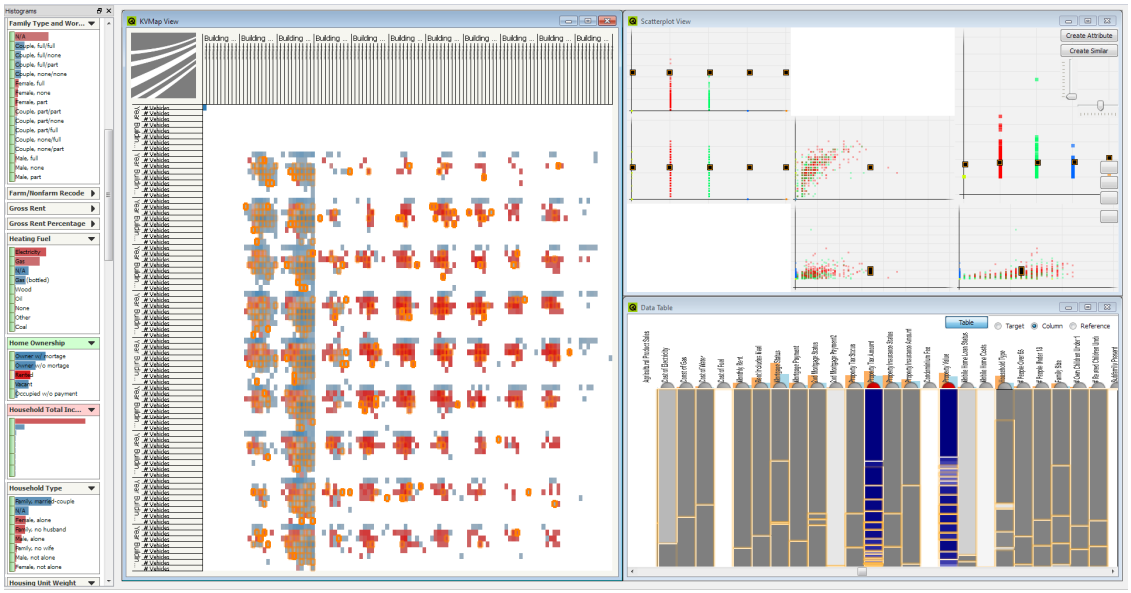


Abbildung 4.23: *Hierbei handelt es sich um das Gesamtsystem, das im Rahmen dieser Arbeit entstanden ist. Die KVMap-Technik, die Smartstripes-Technik, sowie einige weitere bereits existierende Visualisierungen und Data-Mining Verfahren sind in diesem System kombiniert.*

Die Visualisierung für die interaktive Attributauswahl, SmartStripes, beschreibt eine weitere Kopplung zwischen analytischen Verfahren. Wieder sind die Visualisierung und die Interaktion ein Ansatzpunkt dieser Kopplung. Durch den Mixed-Initiative Mode wurde hier eine Verschränkung zwischen Interaktion und Automatismus auf der Heuristik der Verfahren erreicht. Die Freiheitsgrade der Heuristik bei den Verfahren der Attributselektion ist die Menge der gewählten Attribute. Durch die Visualisierung erhält der Anwender Zugriff auf alle Freiheitsgrade dieser Heuristik, und auf Details der Qualitätsmaße zur Bewertung von

Alternativen. Für die weiterführende Forschung interessant ist die Frage, ob diese Verschränkung für beliebige Heuristiken prinzipiell und sinnvoll möglich wäre.
Durch Smartstripes werden Parameter für die Steuerung weiterer Analyseverfahren definiert. In einem System wurden dafür prototypisch Smartstripes, die KVMap Technik und weitere Analyseverfahren kombiniert. Smartstripes liefert hier den Überblick über Zusammenhange zwischen den Attributen, die vor der Analyse bekannt waren und auch den Attributen und Mustern, die während der Analyse erst erzeugt wurden. In diesem System wird dadurch eine Kopplung zwischen detaillierter Analyse - mit frei wählbaren Verfahren - und der überblicksartigen Analyse mit Smartstripes hergestellt. Durch diese Aufteilung wird es möglich, potentiell wertvolle Beziehungen in hochdimensionalen Daten im Fokus zu behalten. Dies gelingt auch dann, wenn die Daten von vornherein sehr viele Attribute enthalten und auch dann noch, wenn während der Analyse zahlreiche neue Zusammenhänge und synthetische Attribute erzeugt wurden.

Kapitel 5

Zusammenfassung und Ausblick

Das hier vorgestellte Konzept behandelt einen Teilschritt in der Datenanalyse - die Suche von Mustern in den Daten und deren Transformation in symbolische Modelle. Für diesen Teilschritt können sowohl visuell-interaktive Techniken aus der Informationsvisualisierung, als auch automatische Techniken aus dem Data-Mining eingesetzt werden. Dabei werden in bisherigen Ansätzen Suche und Modellierung entweder beide durch den Menschen oder beide durch die Maschine durchgeführt.
Dieses Konzept teilt die Aufgaben von Mensch und Maschine dabei neu auf. Der Mensch bringt dabei seine Fähigkeit für die Erkennung von Mustern ein; automatische Verfahren werden für die Modellierung der identifizierten Muster eingesetzt. Die Dimensionalität der Muster, die von Menschen nicht nur gesehen, sondern effektiv genutzt werden können, steigt von zwei bis drei auf zehn Dimensionen und mehr. Spezifische Schwächen von automatischen und visuell-interaktiven Verfahren können bei dieser Aufteilung durch die jeweils komplementäre Technologie größtenteils kompensiert werden:

Das Konzept dieser Arbeit verbessert die Nutzbarkeit von Visualisierungen für hochdimensionale Daten durch die Kopplung visuell-interaktiver und automatischer Analysetechniken. Je mehr Dimensionen die dargestellten Zusammenhänge umfassen, desto weniger ist der Mensch allein in der Lage, deren Komplexität nicht nur visuell, sondern vor allem kognitiv zu erfassen und zu interpretieren (siehe 2.4.4 *Visuelle Wahrnehmung*). Ohne Hilfsmittel oder langes Training ist schon die Interpretation von drei- und mehrdimensionalen Zusammenhängen selbst aus einer an sich geeigneten Visualisierung mühsam.
Dieses Konzept nutzt automatische Verfahren als Hilfsmittel vor der Interpretation komplexerer Beziehungen in den Daten. Die durch den Menschen visuell erfassten Muster werden als Eingabedaten für die automatischen Verfahren genutzt. Dafür muss der Mensch lediglich Teile des Musters durch direkte Selektion auswählen (siehe 2.4.3.1 *Direkte Selektion*). Die automatischen Verfahren transformieren dieses Muster in ein symbolisches Modell. Die auf diese Weise interpretierbaren Muster können wesentlich komplexere Zusammenhänge - mit zehn und mehr Attributen - repräsentieren. Die Interpretation visualisierter Daten wird daher nicht mehr durch das Arbeitsgedächtnis des Menschen auf einfache Beziehungen beschränkt.
Diese Kopplung eröffnet neue Möglichkeiten für das Design von Visualisierungen für die explorative Analyse: Ohne Hilfsmittel muss eine Visualisierungstechnik die Anforderungen

erfüllen, sowohl die *Erkennung* als auch die *Interpretation* von Mustern zu ermöglichen. Dabei sind diese Aufgaben so unterschiedlich, dass man nicht davon ausgehen kann, dass eine Verbesserung der einen Aufgabe stets auch eine Verbesserung der anderen nach sich zieht. Mit der automatischen Transformation als Hilfsmittel ist es dagegen möglich, dedizierte Visualisierungen zu entwickeln, durch die die menschliche Mustererkennung optimal unterstützt werden kann (siehe 3.1 *Separation von Mustererkennung und Musterbeschreibung*).
Das Ergebnis der Modellierung wird ebenfalls durch eine Visualisierung dargestellt. Zur Visualisierung der Daten gehört daher eine komplementäre, interaktive Visualisierung des Modells. Beide Visualisierungen sind über die automatischen Verfahren miteinander gekoppelt, und jede Editierung der Muster oder des Modells hat entsprechende Änderungen in der jeweils anderen Visualisierung zur Folge (siehe 3.1.5 *Modellvisualisierung*). Diese Methode verallgemeinert das in der Informationsvisualisierung als *Brushing & Linking* bekannte Konzept hin zur Kopplung von Visualisierung auf verschiedenen Abstraktionsebenen (siehe 3.2.2 *Visuelles Feedback*).

In dieser Arbeit wurde auch gezeigt, wie diese Kopplung auch die Nutzbarkeit automatischer Verfahren verbessert. Das menschliche Wahrnehmungssystem ist vor allem durch seine Flexibilität eine wertvolle Ergänzung zu automatischen Analysetechniken, denn es kann potentiell interessante Zusammenhänge finden, ohne dass vorher formal definiert werden muss, was gesucht wird. Die Suche nach Mustern ist gerade deshalb der Schwachpunkte automatischer, explorativer Verfahren, weil a-priori festgelegt werden muss, wodurch sich ein Muster auszeichnet (siehe 2.3.3 *Allgemeine Charakterisierung von Data-Mining Verfahren*).
Könnte man garantieren, dass diese Wahl der Verfahren und Parameter immer optimale Ergebnisse liefert, dann könnte man Visualisierungstechniken zu Recht aus diesem Teil der Analyse ausschließen. Eine solche Garantie gibt es im allgemeinen jedoch nicht. Stattdessen muss man sicherstellen, dass man herausfindet, wenn diese Festlegung ungeeignet ist. Zum Konzept gehört daher ein visueller Abgleich zwischen den Mustern, die sich in den Daten manifestieren, und den Mustern, die durch das automatische Verfahren repräsentiert werden können.
Für den visuellen Abgleich wird die Analyse im Prinzip umgekehrt. Die aus den wahrgenommenen Mustern konstruierten Modelle werden dabei wieder in Muster umgewandelt und als Feedback in der gleichen Visualisierung zusammen mit den ursprünglichen Daten dargestellt. Da bei der Modellierung die Daten nicht perfekt reproduziert werden (bzw. werden sollen), kann der Modellierungsfehler beim visuellen Abgleich sichtbar gemacht werden. Jede interaktive Veränderung im Modell und im selektierten Muster wirkt sich auf diesen Abgleich aus.
Durch den visuellen Abgleich kann der Mensch bewerten, ob ein Verfahren prinzipiell geeignet ist, die Muster in den Daten zu modellieren. Ist dies der Fall, kann der Mensch durch die interaktive Vorgabe des Musters festlegen, welche Daten zum Muster gehören sollen und welche nicht. Zusätzlich kann man durch den Abgleich den Modellierungsfehler qualitativ bewerten. Während die quantitativen Gütemaße die Modellierungsfehler auf wenige Kennzahlen aggregieren, erlaubt der visuelle Abgleich eine qualitative Untersuchung des Modells und des Verfahrens.
Der visuelle Abgleich nutzt dabei lediglich die Fähigkeit der Mustererkennung; um zu entscheiden *ob* die Modellierungsfehler lediglich ein Rauschen oder selbst wieder ein überlagertes Muster dargestellen, ist eine Interpretation der Muster beispielsweise nicht notwendig.

Grundsätzlich ist das vorgestellte Konzept unabhängig von bestimmten Visualisierungstechniken und von prädiktiven Data-Mining Verfahren. Dies ist vor allem deshalb notwendig, da ungeeignete Verfahren ausgetauscht werden müssen. Die Wahl des Data-Mining Verfahrens definiert dabei die Art der Modelle, mit denen die Muster beschrieben werden sollen. Die Parameter, die das Verfahren steuern, müssen dabei vorgegeben werden. Beispielhaft wurde jedoch auch gezeigt werden, wie ein numerischer Parameter für die Modellierung innerhalb der Visualisierung der Daten dargestellt und durch die Auswahl des Musters gesteuert werden kann (siehe 3.2.3.1 *Iteratives Feedback*).

Zusammengenommen beschreibt das Konzept zwei Verbindungen zwischen visuell-interaktiven Techniken und automatischen Techniken in entgegengesetzte Richtungen. Mustererkennung, Interaktion und automatische Modellierung beschreiben den Weg von Muster zum Modell (siehe 3.1 *Separation von Mustererkennung und Musterbeschreibung*). Simulation, Feedback und visueller Abgleich beschreiben den Weg vom Modell zurück zum Muster (siehe 3.2 *Konfirmatives Feedback*). Insgesamt wird ein zyklischer, iterativer Prozess etabliert, in dem explorative und konfirmative Analyse miteinander verbunden werden (siehe 3.3 *Synthese*). Durch die zyklische Verkettung der Prozesse wird ein visueller Abgleich analytischer Artefakte auf verschiedenen Abstraktionsebenen etabliert. In diesem Prozess werden im Erfolgsfall die Muster der ursprünglichen Daten und die Muster, die das Modell repräsentiert, konvergieren. Ist dies nicht der Fall, dann wird in der Visualisierung sichtbar, dass das automatische Verfahren nur auf einen Teil der Daten oder sogar überhaupt nicht angewendet werden kann (siehe 3.2.3.1 *Iteratives Feedback*).
Die Bewertung eines Ergebnisses der Modellierung wird in diesem zyklischen Prozess durch eine visuelle Bewertung in einer oder mehrerer Visualisierungen ergänzt. Zusammengefasst ergeben sich aus der neuen Aufgabenverteilung zwischen Mensch und Maschine und der Kombination von explorativer und konfirmativer Analyse folgende Verbesserungen:

- Der wichtigste Aspekt ist, dass der Nutzwert einer Visualisierung hochdimensionaler Daten von den Beschränkungen durch die notwendige Interpretation der Muster durch den Menschen entkoppelt wird. Eine Darstellung von drei- und mehrdimensionalen Zusammenhängen war bisher zwar möglich; sie effektiv zu nutzen erforderte jedoch einen hohen kognitiven Aufwand. Durch die Entlastung des Menschen von der Konstruktion des Modells wird die Komplexität der Darstellung nur noch durch die Fähigkeiten der Mustererkennung begrenzt. Dabei konnte in der Realisierung gezeigt werden, dass auch eine Darstellung von mehr als zehndimensionen Zusammenhängen nutzbar ist.

- Die Möglichkeit der Darstellung komplexerer Zusammenhänge erhöht die Sicherheit bei der Datenanalyse, da diese exponiert und genutzt werden können. Unabhängig davon, wo Visualisierungen innerhalb einer Analyse eingesetzt werden, stellen sie ein „Sichtfenster“ für die Kontrolle von ansonsten verborgenen (Zwischen-)Ergebnissen und Prozessen dar. Dieses Sichtfenster ist immer begrenzt. Man muss davon ausgehen, dass nur solche Artefakte einen Einfluss auf Entscheidungen des Analysierenden haben, die überhaupt exponiert werden. Je mehr Faktoren bei dieser Kontrolle einander in Bezug gesetzt werden können, desto geringer wird die Wahrscheinlichkeit, dass potentiell relevante Zusammenhänge übersehen werden.

- Der visuelle Abgleich erlaubt nicht nur Rückschlüsse auf die Qualität des konstruierten Modells, sondern auch auf die prinzipielle Eignung eines automatischen Verfahrens für dessen Modellierung. Aus den quantitativen Gütemaßen allein ist die Eignung eines Verfahrens nur indirekt ableitbar, weil die Gütemaße ebenso zur Disposition stehen können wie das Verfahren. Durch die Hintereinanderausführung der beiden komplementären Prozesse können die Verfahren gegeneinander validiert werden. Der visuelle Abgleich liefert ein zusätzliches Kriterium für die Eignung der Verfahren, und erhöht so die Sicherheit bei deren Auswahl.

- Im Vergleich zu Methoden, in denen der Mensch die Konstruktion seiner Modelle vornimmt, ist die automatische Modellierung der Muster reproduzierbar. Das Modell ist lokal optimal bezüglich der dafür jeweils spezifizierten Gütekriterien für das Verfahren.

- Im Vergleich zu Methoden, in denen die Suche nach Mustern durch automatische Verfahren durchgeführt wird, ist die Identifizierung von Mustern durch den Menschen robuster. Dies erleichtert die Trennung von Muster und Rauschen, aber auch die Separation mehrerer sich überlagernder Strukturen wird auf diese Weise möglich. Insbesondere wird dadurch der Fall abgedeckt, in dem verschiedene Strukturen nur durch verschiedene Modelle gut dargestellt werden können.

- Durch die Verallgemeinerung des Konzeptes des *Brushing & Linking* können zwei Visualisierungen nicht nur auf der gleichen Abstraktionsebene miteinander verbunden werden (als verschiedene Perspektiven auf die Daten). Zwei Visualisierungen können auch unterschiedlichen Abstraktionsebenen der Analyse miteinander in Korrespondenz setzen (als Verbindung von Daten und Modell).

Diese Arbeit gründet auf mehreren Modellen aus verschiedenen Forschungsbereichen, die teilweise detailliert, teilweise verallgemeinert werden:
Das Visual-Analytics-Prozessmodell (siehe 2.5.1) illustriert dabei die Idee der direkten, technischen Kopplung zwischen Techniken der Informationsvisualisierung und des Data-Mining - im Gegensatz zu einer methodischen Kopplung. Dieses Modell beschreibt jedoch nicht, auf welche Weise die Technologien miteinander verknüpft werden. Aus diesem Grund wurde dieses Modell verfeinert, um die beiden hier vorgestellten Kopplungsvarianten zwischen visuell-interaktiven und automatischen Verfahren von anderen zu unterscheiden und abzugrenzen. Bei dieser Systematisierung wurden acht verschiedene technische Kopplungsvarianten vorgestellt und existierenden Ansätzen zugeordnet.
Die Kriterien für die Systematisierung gründen auf den älteren Modellen für den Knowledge-Discovery-Process (siehe 2.2) und die Informationsvisualisierung (siehe 2.4.2). Diese Modelle wurden konsultiert, um dem Anspruch einer allgemeinen Einordnung verschiedener Techniken gerecht zu werden. Beide Modelle beschreiben eine Datenverarbeitungspipeline. Auf der Basis dieser Modelle konnten kompatible, allgemeine Ansatzpunkte für eine Kopplung identifiziert werden: Modellparameter, Verfahrensparameter und Daten. In der Systematik wird eine Kopplung nach diesen Ansatzpunkten und der Richtung des Informationsflusses charakterisiert. Ein Ergebnis der Zuordnung ist, dass es zu jeder Kopplungsvariante beispielhafte Ansätze gibt, diese jedoch insgesamt ungleich häufig erforscht werden: So wird eine

Visualisierung als Daten*quelle*, wie im hier vorgestellten Konzept, vergleichsweise selten eingesetzt - der im Konzept vorgestellte zyklische Prozess von Mustererkennung, Interaktion, Modellierung und Feedback ist neu.
Über die Abgrenzung der eigenen Arbeit hinaus soll die Systematisierung auch dazu dienen, die technischen Möglichkeiten der Verbindung dieser Technologien - unabhängig von spezifischen Techniken - zu beschreiben und verdeutlichen. So konnten Lücken identifiziert werden, die in weiterführenden Arbeiten erforscht werden sollen. Beispielsweise wird die Nutzung von Visualisierungstechniken für die Steuerung der Verfahrensparameter automatischer Techniken seltener beforscht, wenngleich Visualisierungstechniken in den Ingenieurwissenschaften durchaus schon für die Steuerung von Prozessen genutzt werden. Diese Kopplungsvariante liegt nicht im Fokus dieser Arbeit und wurde daher nur an einem Beispiel illustriert (siehe 4.2.2 *Unabhängigkeitstests für die Vorauswahl der Attribute*). Durch Erforschung dieser Verbindung könnte die Visualisierung und die Prüfung verschiedener Verfahenskonfigurationen noch häufiger die methodische Grundlage für die sicherere Parametrisierung und Automatisierung der Analyseprozesse liefern.
Der Vergleich des Knowledge-Discovery-Prozesses and des Informationsvisualisierungsprozesses führte viele prinzipielle Gemeinsamkeiten zutage, die im Rahmen dieser Arbeit eigentlich nur punktuell betrachtet werden können. Allgemeiner betrachtet ergibt sich die Fragestellung nach den technischen und methodischen Vorraussetzungen für den Austausch kompatibler Techniken aus beiden Technologien in den jeweiligen Pipelines (siehe 5.1).

Die Trennung zwischen Mustererkennung und Modellierung gründet auf den Modellen für die Interaktion (siehe 2.4.3) und der Wahrnehmung (siehe 2.4.4). In beiden Modellen werden diese beiden Aufgaben bereits unterschieden. Unter Berufung auf das Wahrnehmungsmodell von Ware kann man argumentieren, dass beide Aufgaben einen unterschiedlichen kognitiven Aufwand erfordern und durch unterschiedliche Methoden und Designs erleichtert und auch behindert werden können.
Im technischen Modell des Informationsvisualisierungsprozess (s.o.) wird eine solche Unterscheidung nicht gemacht. Beide Aufgaben müssen - unter potentiell unterschiedlichen Voraussetzungen - durch den Menschen durchgeführt werden. Der Aufwand für die Nutzung einer Technik ist dann bestimmt durch jene Aufgabe, die schwerer fällt bzw. jene, die schlechter unterstützt wird. Für eine einseitige Verbesserung eines Designs bestünde kein Bedarf - selbst wenn sie nicht zu Lasten der anderen Aufgabe ginge.
Dagegen ist es durchaus möglich, komplexere Beziehungen darzustellen und wahrzunehmen als ein Mensch direkt beschreiben kann. Wenn daher das Design einer Visualisierung keine Kompromisslösung zwischen Wahrnehmbarkeit und Lesbarkeit darstellen muss, können die Fähigkeiten des Menschen, Muster und Strukturen zu erkennen, besser unterstützt werden. Die Modellierung durch die automatischen Verfahren entlastet den Nutzer und erlaubt eine größere Spezialisierung beim Design von Visualisierungstechniken.
Durch die Entlastung von der Modellierungsaufgabe verändern sich die Modelle für die Wahrnehmung und die Interaktion (siehe 3.1.2 *Erweiterung des Visual-Analytics-Prozess*), da die Erzeugung eines Modells aus dem Muster durch Verfahren des Data-Mining automatisch durchgeführt wird. Dabei operieren die automatischen Verfahren nicht direkt auf den Daten, sondern auf einer Datensegmentierung, die der Mensch in der Interaktion als Muster identifiziert hat. Die Aufgabe der automatischen Verfahren ist daher nicht die *Suche* nach Mustern und Strukturen, sondern die Beschreibung von durch den Menschen vorgegebenen

Strukturen im Suchraum aller Modellparameter. Da dadurch die Anzahl der Freiheitsgrade reduziert wird, vereinfacht sich die Optimierungsaufgabe für die Modellierung. Zudem erlaubt die interaktive, manuelle Segmentierung eine Anpassung der Modelle an Teilstrukturen eines Datensatzes selbst dann, wenn das Verfahren Strukturen des ganzen Datensatzes nur ungenügend beschreiben kann.
Die Trennung zwischen Mustererkennung und Musterbeschreibung ist für die Analyse komplexer Muster nur dann sinnvoll, wenn der Nutzer das Korrespondenzproblem zwischen Wahrnehmung und Interaktion nicht selbst lösen muss. Dies schränkt die Möglichkeiten der Interaktion erheblich ein. Die für die Interaktion identifizierten Prinzipien lassen sich darin zusammenfassen, dass das Korrespondenzproblem entweder trivial ist, oder dass die Korrespondenz durch den Menschen bereits gelernt wurde (siehe 2.4.3.1). Ausgehend von dieser Einschränkung wurden die Daten identifiziert, die durch die Interaktion als Datenquelle für die Kopplung dienen können.

Direkte Selektion ist in der Informationsvisualisierung häufig die Grundlage für komplexere Interaktion. Für dieses Konzept am wichtigsten ist hierbei die Idee des *Brushing & Linking* - die Kopplung der Interaktion in mehreren Visualisierungstechniken der gleichen Daten (siehe 2.4.3.2). Wird in einer Visualisierung eine Menge von Daten selektiert und hervorgehoben, so wird in der oder den anderen Visualisierungen die korrespondierende Menge ebenfalls hervorgehoben. Das Datenmodell wird dabei als gemeinsame Referenz genutzt.
Das hier vorgestellte Konzept kann daher auch als eine Verallgemeinerung des *Brushing & Linking* aufgefasst werden. Anstatt die Korrespondenz zwischen den Visualisierungen innerhalb des Datenmodells herzustellen, wird eine Korrespondenz zwischen dem Datenmodell und einem prädiktiven analytischen Modell gesucht. Durch diese Erweiterung des *Brushing & Linking* kann eine Visualisierung der Daten und eine Visualisierung des analytischen Modells (siehe 3.2.3.1 *Iteratives Feedback*) miteinander verbunden werden. Durch eine Interaktion in der Visualisierung der Daten, und eine Interaktion in der Visualisierung des Modells wird die Darstellung auf der jeweils anderen Abstraktionsebene angepasst. Der zyklische Prozess (Mustererkennung - Interaktion - Modellierung - Feedback) wird dabei im Prinzip an unterschiedlichen Stellen begonnen.

Das hier vorgestellte Konzept wurde prototypisch umgesetzt. Der Prototyp besteht aus vier Komponenten. Die erste Komponente ist eine Visualisierungstechnik für mehrdimensionale Daten, die gleichzeitig auch in der Lage ist, mehrdimensionale Bezüge zwischen diesen Daten darzustellen (siehe 4.1 *KVMap*). Die Komplexität der potentiell darstellbaren - und auch Erkennbaren - Zusammenhänge war der Ansatzpunkt für die anderen Teile des Prototypen: Die Kopplung der Visualisierungstechniken mit einer Technik für die automatischen Datenanalyse - in diesem Fall Entscheidungsbäume für die Modellierung (siehe Abschnitt 4.1.4). Die dritte Komponente dient der Visualisierung und auch der Editierung des formalen Modells (d.h. des Entscheidungsbaums) (siehe 4.1.4.4). Wie die automatische Modellierung ist auch die interaktive Veränderung des Modells mit der Visualisierung der Daten gekoppelt.
Die vierte Komponente ist eine Visualisierung, die der Auswahl von Attributen aus einer hochdimensionalen Datentabelle dient (siehe Abschnitt 4.2.2). Diese Auswahl ist ein vorbereitender Schritt zur eigentlichen Analyse mit den vorgenannten Komponenten. Die Verbindung setzt eine visuell-interaktive Steuerung von Verfahrensparametern - den gewählten Attributen - prototypisch um.

5.1 Ausblick - Metaanalyse

In dieser Arbeit wurde ein Ansatz vorgestellt, der es einerseits erlaubt, die Qualität eines analytischen Modells zu bewerten, und der es andererseits erlaubt, die Qualität des Verfahrens zu bewerten, mit dem dieses Modell konstruiert wird. Die Frage, ob ein Verfahren überhaupt geeignet ist, eine Korrespondenz zwischen Rohdaten und formalen Modell zu beschreiben, ist eine Problemstellung der Metaanalyse.
Diese Problemstellung lässt sich allgemeiner fassen: Wie in Abschnitt 2.1.1 beschrieben, müssen Entscheidungen über Verfahren, Parameter, etc. in der Analyse teilweise unter unsicherem Wissen getroffen werden. Diese Entscheidungen haben mittelbar oder unmittelbar Einfluss auf der Ergebnis. Die Voraussetzung für das Ziel der Analyse - der Elimination von Unsicherheit für den Anwender bezüglich der Fragestellungen der Anwendungsdomäne - ist daher die Elimination von Unsicherheit bezüglich jeder Entscheidung *während* der Analyse.
Für die Zukunft stellt sich zunächst die Frage, ob eine Fragestellung einer Anwendungsdomäne, die eine Analyse letztlich motiviert, sich grundsätzlich von einer Fragestellung unterscheidet, nach der eine Entscheidung innerhalb der Analyse getroffen werden muss. Sollte dies nicht der Fall sein, besteht die Möglichkeit, die Entscheidungen des Analysten mit dem gleichen Repertoire an Verfahren zu untersuchen, das für die Analyse generell zur Verfügung steht. Dazu zählen dementsprechend auch Visualisierungstechniken ebenso wie automatische Verfahren. Über die Verbindung von Analyse und Metaanalyse bieten sich weitere Möglichkeiten für die Kopplung.
Ein Ziel der Metaanalyse besteht darin, sowohl die Optionen der Entscheidungen eines Analysten zu exponieren, als auch die Bezüge zwischen den Entscheidungen, der Fragestellung und den Ergebnissen der Analyse sichtbar zu machen. Ein zweites Ziel besteht darin, verschiedene Entscheidungsoptionen der Analyse zu untersuchen und deren Effekte miteinander zu vergleichen. Langfristig könnten Visualisierungstechniken dabei die Funktion übernehmen, Unsicherheit und methodische Schwächen in der Analyse so augenfällig darzustellen, dass es unmöglich wird, sie zu ignorieren. Dass dies auch bei komplexen Entscheidungen, wie der Wahl eines automatischen Analyseverfahrens nicht unmöglich ist, wurde im Konzept bereits gezeigt.

5.2 Ausblick - Konfirmative Analyse für deskriptive Modelle

Im Konzept wurde ausgenutzt, dass prädiktive Modelle eine funktionale Abhängigkeit zwischen Attributen des gleichen Datenobjekts beschreiben. Die konfirmative Analyse eines prädiktiven Verfahrens ist deshalb vergleichsweise einfach, weil potentielle Abhängigkeiten zwischen verschiedenen Datensätzen ignoriert werden können. Der Kontext, innerhalb dessen der Abgleich zwischen Ergebnis und Referenz stattfindet, ist ein Datenobjekt und alle Daten, die diesem Objekt eindeutig zuzuordnen sind. Die funktionale Abhängigkeit kann für die Überprüfung eines prädiktiven Verfahrens genutzt werden, weil sie die Simulation (im Sinne einer Vorhersage) explizit definiert. Die Transformation zwischen Modell und Daten

kann überdies unabhängig von dem Verfahren bestimmt werden, mit dem das Modell definiert wurde.
Im Gegensatz dazu ist der Kontext innerhalb dessen eine konfirmative Analyse für deskriptive Verfahren stattfinden kann, nicht so eindeutig einzugrenzen. Die Unterscheidung zwischen einer „richtigen“ oder „falschen“ Beschreibung lässt sich nicht direkt durch einen „Groundtruth“-Vergleich treffen. Im Unterschied zu prädiktiven Verfahren konstruiert ein deskriptives Verfahren Aussagen über eine Menge von Datenobjekten, die auf wenige charakteristische Deskriptoren reduziert werden. Die Beschreibung ist infolgedessen abhängig von allen Elementen dieser Menge.
Zu den wichtigsten Charakteristika in der Datenanalyse gehören die Qualitätskennzahlen, mit denen die Güte *prädiktiver* Verfahren beschrieben wird; in einer (binären) Klassifikation sind dies beispielsweise etwa die Wahrscheinlichkeiten für „falsch-positive“ oder „falsch-negative“ Ergebnisse des Prädiktors. In diesem Fällen wird eine ganze Teilmenge des Datensatzes, auf jeweils ein numerisches Merkmal reduziert.
An dieser Stelle muss unterschieden werden zwischen der konfirmativen Analyse *durch Anwendung* deskriptiver Verfahren und der konfirmativen Analyse *für die Bewertung* deskriptiver Verfahren. Im ersten Fall werden die Deskriptoren einer bestimmten Menge von Datensätzen berechnet, so dass sie mit Referenzdeskriptoren verglichen werden können. Die Verfahren, die die Beschreibung konstruieren werden durch die konfirmative Analyse weder in Frage gestellt noch verändert. Ein Beispiel dafür sind statistische Kennzahlen wie Mittelwert und Median, praktisch jede Art von Qualitätskennzahlen oder auch die Aggregation, mit der im letzten Kapitel alle Datenobjekte zusammengefasst wurden, die das gleiche Merkmalsprofil aufwiesen (siehe Abschnitt 4.1.3.3).
Eine andere Perspektive bietet jedoch der zweite Fall, hinter dem sich die Fragestellung verbirgt, ob die in der Beschreibung zusammengefassten Informationen relevant sind für die Fragestellung, die mit der Analyse verfolgt werden soll. Ebenso kritisch ist die Frage, ob durch die Zusammenfassung so viel Information verloren geht, dass die Bedeutung der Informationen in der Beschreibung geringer ist als die Bedeutung der Informationen in den beschriebenen Daten.
Im Rahmen des Konzepts wurde ausgenutzt, dass die Konstruktion eines prädiktiven Modells durch die Simulation einen komplementären Prozess besitzt, der eine gegenseitige Validierung erlaubt. Als Ausblick stellt sich hier die Frage, ob es eine analoge Strategie für deskriptive Modelle gibt. Dabei würden komplementäre, unabhängige Transformationen eine Korrespondenz zwischen Daten und Modellen herstellen, und die Validierung würde durch einen visuellen Abgleich auf jeder Abstraktionsebene durchgeführt.
Der Vergleich ist eine fundamentale Operation der Analyse, die eine Beschreibung überhaupt motiviert. Bei deskriptiven Modellen geht es um den Vergleich von Mengen (anstatt um den Vergleich von Werten). Die Qualität der durch die Beschreibung erzeugten Informationen hängt ab vom Beschreibungsmodell, aber auch davon, ob die Mengen eine Bedeutung jenseits dieses Modells haben. Ein an sich geeignetes Beschreibungsmodell ist wertlos, wenn die Mengen, auf die diese Beschreibung angewendet wird, irrelevant sind. Umgekehrt sind Mengen, die die Extension potentiell sinnvoller Konzepte darstellen, wertlos, wenn das Beschreibungsmodell ungeeignet ist. Ziel der deskriptiven Analyse muss es sein, gute Mengen *und* gute Beschreibungen zu finden. Dies funktioniert im allgemeinen nicht, wenn man beides a-priori festlegt.

Allgemein kann man für die Wahl eines deskriptiven Verfahrens mehrere Qualitätskriterien angeben, die jeweils unterschiedliche Modelle zueinander in Bezug setzen: Ein Kriterium ist die Stärke des Bezugs der Deskriptoren zur Fragestellung der Analyse. Nur dann, wenn die Deskriptoren Informationen repräsentieren, die mittelbaren oder unmittelbaren Einfluss auf Entscheidungen haben oder die Grundlage für die Bildung von ihrerseits relevanten Begriffe sein können, ist die Beschreibung sinnvoll.
Ein weiteres Kriterium ist, ob und in welcher Qualität das Muster aus seiner Beschreibung reproduziert werden kann. Wie bei prädiktiven Verfahren muss durch die Vereinfachung der Beschreibung auch ein Fehler in Kauf genommen werden. Will man die Fragestellung der konfirmativen Analyse in der Form, wie sie im Rahmen dieses Konzepts für prädiktive Modelle gelöst wurde, auf deskriptive Modelle übertragen, muss geklärt werden, wie ein visueller Abgleich in den Daten ermöglicht werden kann. Dieses Kriterium beschreibt daher den Bezug des Beschreibungsmodells zum Datenmodell des Merkmalsraums.
Da viele verschiedene Muster auf die gleiche Beschreibung abgebildet werden, ist ein drittes Kriterium, ob die Mengen, die auf die gleichen Deskriptoren abgebildet, auch bezüglich anderer Kriterien ähnlich sind, die unabhängig sind vom Beschreibungsmodell.

5.3 Schlusswort

Beschreibt man das Konzept in allgemeiner Form, handelt es sich um Methoden, mit denen die Korrespondenz zwischen verschiedenen Modellen hergestellt, beschrieben und überprüft wird. Dabei handelt es sich bei den Modellen sowohl um Datenmodelle als auch um analytische Modelle und die visuelle Repräsentierung. Die Korrespondenz zwischen diesen Modellen wird dabei durch automatische Verfahren formalisiert. Der visuelle Abgleich auf der Ebene der Daten und der Ebene der Modelle verbindet dabei die Überprüfung der Modelle und die der Verfahren, die sie konstruieren.
Der Gedanke, der hinter der Nutzung von Visualisierung in der Analyse steht, ist nicht nur die Darstellung von Ergebnissen, sondern das Potential, fehlerhafte Entscheidungen in der Analyse auf einen möglichst einfachen Vergleich zurückzuführen, der so weit wie möglich unabhängig von formalen Kriterien ist. Fehler können auf diese Weise wie Muster identifiziert werden, ohne dass für sie eine formale Beschreibung vorliegt.
Wie in der Einleitung beschrieben, war der Ausgangspunkt der hier vorgestellten Überlegungen eine Visualisierungstechnik, die zwar hochdimensionale Bezüge zwischen den Daten sichtbar machte, die Muster jedoch so komplex waren, dass ein Mensch sie ohne Hilfsmittel nicht beschreiben kann. Diese Visualisierungstechnik wurde im vorherigen Kapitel vorgestellt. Die Möglichkeit, Zusammenhänge so sichtbar zu machen, dass sie ihrer tatsächlichen Komplexität gerecht werden, ist unabhängig davon relevant, ob sie sofort beschrieben und erklärt werden können. Denn es kann bereits als Fortschritt gewertet werden, wenn bei der Visualisierung von Ergebnissen dort weniger Sicherheit vermittelt wird, wo überhaupt keine Sicherheit vorliegt - wenn komplexe Wahrheiten nicht von einfachen oder gar gewünschten Wahrheiten maskiert werden können. Jede Information, die den Anstoß dafür geben könnte, die Verfahren und Ergebnisse einer Analyse einer Revision zu unterziehen, verbessert letztlich Analyse und Entscheidungen.
Denkt man diese Idee weiter, stellt sich die Frage, ob eine Analyse jemals so exponiert wer-

den kann, dass solche Informationen auch den betroffenen Laien (d.h. Nicht-Analysten) eine kritische Bewertung der Konsistenz von Verfahren und Ergebnissen ermöglichen. Auch wenn keine formale Ausbildung notwendig ist, um Bilder zu vergleichen, ist dies heute noch nicht abzusehen. Eine Exponierung analytischer Ergebnisse in einer Form, die einen Menschen anregt, sich mit der Komplexität der Zusammenhänge zu konfrontieren, anstatt ihr auszuweichen, erscheint dennoch ein lohnendes Ziel.
Weiterhin gilt, dass jede einzelne Visualisierung nur einen Ausschnitt aller Zusammenhänge innerhalb der Daten sichtbar machen kann, wobei die Wahl der Visualisierung die Art der Zusammenhänge determiniert. Da das hier vorgestellte Konzept die Herstellung einer Korrespondenz zwischen empirischen Daten und Modellen beschreibt, ist die Kombination von Techniken für die Konsolidierung mehrerer verschiedener Korrespondenzen der folgende, nahezu zwingende Schritt. In den beiden Thematiken, die im Ausblick angesprochen wurden, wurde die Möglichkeit angedeutet, dass das gleiche Repertoire von Methoden und Techniken auch für diese Aufgabe genutzt werden kann. Die damit verbundenen Fragestellungen bieten zahlreiche interessante Ansatzpunkte für die Forschung in den kommenden Jahren.

Literaturverzeichnis

[AEEK99] Mihael Ankerst, Christian Elsen, Martin Ester, and Hans-Peter Kriegel. Visual classification: an interactive approach to decision tree construction. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 392–396, New York, NY, USA, 1999. ACM.

[AEK00] Mihael Ankerst, Martin Ester, and Hans-Peter Kriegel. Towards an effective cooperation of the user and the computer for classification. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 179–188, New York, NY, USA, 2000. ACM.

[AES05] Robert Amar, James Eagan, and John Stasko. Low-level components of analytic activity in information visualization. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*, page 15, Washington, DC, USA, 2005. IEEE Computer Society.

[AGGR98] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. *SIGMOD Rec.*, 27(2):94–105, 1998.

[And96] J.R. Anderson. *Kognitive Psychologie*. Spektrum Akademischer Verlag, Heidelberg, 1996. 2. Auflage.

[AS94] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *Proceeding of the 20th International Conference on Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994.

[AS05] Robert A. Amar and John T. Stasko. Knowledge precepts for design and evaluation of information visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):432–442, 2005.

[AWS92] Christopher Ahlberg, Christopher Williamson, and Ben Shneiderman. Dynamic queries for information exploration: an implementation and evaluation. In *CHI '92: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 619–626, New York, NY, USA, 1992. ACM.

[BCK08] Shyam Boriah, Varun Chandola, and Vipin Kumar. Similarity measures for categorical data: A comparative evaluation. In *SDM '08: SIAM International Conference on Data-Mining*, pages 243–254, Philadelphia, PA, USA, 2008. SIAM.

[BDM07] Jolita Bernatavičienė, Gintautas Dzemyda, and Virginijus Marcinkevičius. Conditions for optimal efficiency of relative MDS. *Informatica*, 18(2):187–202, 2007.

[Ben06] Fabian Bendix. Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):558–568, 2006.

[Ber83] J. Bertin. *Semiology of Graphics*. University of Wisconsin Press, Madison, WI, 1983. (trans. W. Berg).

[Ber02] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[BFOS84] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[BGH+06] Ella Bingham, Aristides Gionis, Niina Haiminen, Heli Hiisilä, Heikki Mannila, and Evimaria Terzi. Segmentation and dimensionality reduction. In *Proceedings of the Sixth SIAM International Conference on Data Mining, April 20-22, 2006, Bethesda, MD, USA*, 2006.

[BGP+11] Michelle Borkin, Krzysztof Gajos, Amanda Peters, Dimitrios Mitsouras, Simone Melchionna, Frank Rybicki, Charles Feldman, and Hanspeter Pfister. Evaluation of artery visualizations for heart disease diagnosis. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2479 –2488, dec. 2011.

[BN01] T. Barlow and P. Neville. Case study: visualization for decision tree analysis in data mining. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*, pages 149–152, Washington, DC, USA, 2001. IEEE Computer Society.

[BPFG11] W. Berger, H. Piringer, P. Filzmoser, and E. Gröller. Uncertainty-aware exploration of continuous parameter spaces using multivariate prediction. *Computer Graphics Forum*, 30(3):911–920, 2011.

[BPKG07] Julien Blanchard, Bruno Pinaud, Pascale Kuntz, and Fabrice Guillet. A 2d-3d visualization support for human-centered rule mining. *Computers & Graphics*, 31(3):350–360, 2007.

[Bre99] Cynthia A. Brewer. Color use guidelines for data representation. In *Proceedings of the Section on Statistical Graphics*, pages 55–60, Baltimore, MD, USA, 1999. American Statistical Association.

[Bre01a] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[Bre01b] Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231, 2001.

[Bro09] Gavin Brown. A new perspective for information theoretic feature selection. In *Journal of Machine Learning Research - Proceedings Track*, volume 5, pages 49–56, 2009.

[BT98] Christopher M. Bishop and Michael E. Tipping. A hierarchical latent variable model for data visualization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(3):281–293, 1998.

[BTK11] Enrico Bertini, Andrada Tatu, and Daniel A. Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2203–2212, December 2011.

[Bur04] Remo Aslak Burkhard. Learning from architects: The difference between knowledge visualization and information visualization. In *Proceedings of the 8th International Conference on Information Visualisation, IV 2004, 14-16 July 2004, London, UK*, pages 519–524, 2004.

[BW08] Sven Bachthaler and Daniel Weiskopf. Continuous scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1428–1435, November - December 2008.

[BZL+08] S. Barlowe, Tianyi Zhang, Yujie Liu, Jing Yang, and D. Jacobs. Multivariate visual explanation for high dimensional datasets. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 147–154, Washington, DC, USA, 2008. IEEE Computer Society.

[CBY10] Yang Chen, Scott Barlowe, and Jing Yang. Click2annotate: Automated insight externalization with rich semantics. In *IEEE Symposium on Visual Analytics Science and Technology (VAST) 2010*, pages 155–162, Washington, DC, USA, 2010. IEEE Computer Society.

[CCAD01] Nathan Chia, Richard Cant, and David Al-Dabass. New anti-aliasing and depth of field techniques for games graphics. In *2nd International Conference on Intelligent Games and Simulation (GAME-ON 2001)*, pages 115–, London, UK, 2001.

[CCK+00] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. CRISP-DM 1.0 step-by-step data mining guide. Technical report, The CRISP-DM consortium, August 2000.

[CHD+03] Ping Chen, Chenyi Hu, Wei Ding, Heloise Lynn, and Yves Simon. Icon-based visualization of large high-dimensional datasets. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM)*, page 505, Washington, DC, USA, 2003. IEEE Computer Society.

[Che95] Hsinchun Chen. Machine learning for information retrieval: neural networks, symbolic learning, and genetic algorithms. *Journal of the American Society for Information Science*, 46(3):194–216, 1995.

[Che04] Hong Chen. Compound brushing explained. *Information Visualization*, 3(2):96–108, 2004.

[Chi00] Ed H. Chi. A taxonomy of visualization techniques using the data state reference model. In *INFOVIS '00: Proceedings of the IEEE Symposium on Information Vizualization 2000*, page 69, Washington, DC, USA, 2000. IEEE Computer Society.

[CL04] Keke Chen and Ling Liu. Vista: validating and refining clusters via visualization. *Information Visualization*, 3(4):257–270, 2004.

[CM04] Monica Crubézy and Mark A. Musen. Ontologies in support of problem solving. In Staab and Studer [SS04b], pages 321–342.

[CMS99] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999.

[CR98] Ed Huai-hsin Chi and John Riedl. An operator interaction framework for visualization systems. In *INFOVIS '98: Proceedings of the 1998 IEEE Symposium on Information Visualization*, pages 63–70, Washington, DC, USA, 1998. IEEE Computer Society.

[CWR06] Qingguang Cui, Matthew O. Ward, and Elke A. Rundensteiner. Enhancing scatterplot matrices for data with ordering or spatial attributes. *Proceedings of the SPIE*, 6060:60600R–60600R–11, 2006.

[DKR07] Mark Derthick, John Kolojejchick, and Steven F. Roth. An interactive visualization environment for data exploration. In *Proceedings of Knowledge Discovery in Databases SIGKDD*, pages 2–9. AAAI Press, 2007.

[dSB04] Selan dos Santos and Ken Brodlie. Gaining understanding of multivariate and multidimensional data through visualization. *Computers & Graphics*, 28(3):311–325, 2004.

[EDF08] Niklas Elmqvist, Pierre Dragicevic, and Jean-Daniel Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics*, 17(6):1141–1148, 2008.

[EKSX96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR*, pages 226–231, 1996.

[EMS97] Floriana Esposito, Donato Malerba, and Giovanni Semeraro. A comparative analysis of methods for pruning decision trees. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(5):476–491, 1997.

[Fai07] Joe Faith. Targeted projection pursuit for interactive exploration of high-dimensional data sets. In *Proceedings of 11th International Conference on Information Visualisation (IV)*, 2007.

[Fek04] Jean-Daniel Fekete. The infovis toolkit. In *INFOVIS '04: Proceedings of the IEEE Symposium on Information Visualization*, pages 167–174, Washington, DC, USA, 2004. IEEE Computer Society.

[Few09] Stephen Few. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, Oakland, CA, USA, 2009.

[Fod02] I.K Fodor. *A survey of dimension reduction techniques*. Technical Report UCRL-ID-148494, Lawrence Livermore National Laboratory, Center for Applied Scientific Computing, 2002.

[FPSS96a] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.

[FPSS96b] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *KDD*, pages 82–88, 1996.

[Fri02] Michael Friendly. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56:316–324, November 2002.

[FSL05] Camilla Forsell, Stefan Seipel, and Mats Lind. Simple 3d glyphs for spatial multivariate data. In *INFOVIS '05: Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, page 16, Washington, DC, USA, 2005. IEEE Computer Society.

[FT74] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, 23(9):881–890, 1974.

[GE03] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Resarch*, 3:1157–1182, 2003.

[GH95] M. Garland and P. Heckbert. Fast polygonal approximation of terrains and height fields. Technical Report CMU-CS-95-181, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, September 1995.

[GNRM08] S. Garg, E. Nam, IV. Ramakrishnan, and K. Mueller. Model-driven visual analytics. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 19–26, Washington, DC, USA, 2008. IEEE Computer Society.

[Gol89] David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[GRM10] S. Garg, IV. Ramakrishnan, and K. Mueller. A visual analytics approach to model learning. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 67–74, Washington, DC, USA, 2010. IEEE Computer Society.

[GS06] Henning Griethe and Heidrun Schumann. The visualization of uncertain data: Methods and problems. In *Proceedings of the Simulation und Visualisierung 2006 (SimVis 2006), 2-3 März, Magdeburg*, pages 143–156. SCS Publishing House e.V., 2006.

[Guo03] Diansheng Guo. Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization*, 2:232–246, 2003.

[GW96] Bernhard Ganter and Rudolf Wille. *Formale Begriffsanalyse - Mathematische Grundlagen*. Springer Verlag, Berlin, Heidelberg, New York, 1996.

[HA04] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[HCL05] Jeffrey Heer, Stuart K. Card, and James A. Landay. prefuse: a toolkit for interactive information visualization. In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 421–430, New York, NY, USA, 2005. ACM.

[HDK+07] Ming C. Hao, Umeshwar Dayal, Daniel A. Keim, Dominik Morent, and Joern Schneidewind. Intelligent visual analytics queries. In *VAST '07: Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*, pages 91–98, Washington, DC, USA, 2007. IEEE Computer Society.

[Hea99] Matti A. Hearst. User interfaces and visualization. In Ricardo Baeza-Yates and Berthier Ribeiro-Neto, editors, *Modern Information Retrieval*, chapter 10. Addison-Wesley Longman Publishing Company, 1999.

[Hec08] David Heckerman. A tutorial on learning with bayesian networks. In Dawn E. Holmes and Lakhmi C. Jain, editors, *Innovations in Bayesian Networks*, volume 156 of *Studies in Computational Intelligence*, pages 33–82. Springer, 2008.

[HFM07] Nathalie Henry, Jean-Daniel Fekete, and Michael J. McGuffin. Nodetrix: a hybrid visualization of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1302–1309, 2007.

[HG93] G.J. Hunter and M.F. Goodchild. Managing uncertainty in spatial databases: Putting theory into practice. *Journal of the Urban and Regional Information Systems Association*, 5(2):55–62, 1993.

[HK00] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

[HMM00] Ivan Herman, Guy Melançon, and M. Scott Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000.

[HNN02] TuBao Ho, TrongDung Nguyen, and DungDuc Nguyen. Visualization support for a user-centered kdd process. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 519–524, New York, NY, USA, 2002. ACM.

[Hof01] Donald D. Hoffman. *Visuelle Intelligenz: Wie die Welt im Kopf entsteht*. Klett-Cotta, Stuttgart, 2001. deutsche Ausgabe.

[Hol06] Danny Holten. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):741–748, 2006.

[HR07] Jeffrey Heer and George Robertson. Animated transitions in statistical data graphics. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1240–1247, 2007.

[HS04] Harry Hochheiser and Ben Shneiderman. Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, 2004.

[HSM01] David J. Hand, Padhraic Smyth, and Heikki Mannila. *Principles of data mining*. MIT Press, Cambridge, MA, USA, 2001.

[HvW08] Danny Holten and Jarke van Wijk. Visual comparison of hierarchically organized data. *Computer Graphics Forum*, 27(3):759–766, 2008.

[Hyv99] Aapo Hyvärinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.

[IMI+10] Stephen Ingram, Tamara Munzner, Veronika Irvine, Melanie Tory, Steven Bergner, and Torsten Möller. Dimstiller: Workflows for dimensional analysis and reduction. In *Proceedings of the 5th IEEE Conference on Visual Analytics in Science and Technology (VAST)*, Washington DC, USA, October 2010. IEEE Computer Society.

[JDM00] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition - a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[JFLC08] J. Johansson, C. Forsell, M. Lind, and M. Cooper. Perceiving patterns in parallel coordinates: determining thresholds for identification of relationships. *Information Visualization*, 7(2):152–162, 2008.

[JJ09] Sara Johansson and Jimmy Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics*, 15:993–1000, November 2009.

[JK03] T. J. Jankun-Kelly. *Visualizing visualization: a model and framework for visualization exploration.* PhD thesis, University of California, Davis, 2003. Adviser-Ma, Kwan-Liu.

[JKMG07] T. J. Jankun-Kelly, Kwan-Liu Ma, and Michael Gertz. A model and framework for visualization exploration. *IEEE Transactions on Visualization and Computer Graphics*, 13(2):357–369, 2007.

[JKP94] G.H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine Learning, New Brunswick, NJ*, pages 121–129. Morgan Kaufmann Publishers, San Francisco, CA, 1994.

[JLJC05] Jimmy Johansson, Patric Ljung, Mikael Jern, and Matthew Cooper. Revealing structure within clustered parallel coordinates displays. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, INFOVIS '05, pages 17–, Washington, DC, USA, 2005. IEEE Computer Society.

[JTS08] Mathias John, Christian Tominski, and Heidrun Schumann. Visual and Analytical Extensions for the Table Lens. In *Proceedings of Visualization and Data Analysis (VDA)*, pages 680907–1–680907–12. SPIE/IS&T, 2008.

[KAF+08] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melancon. Visual analytics: Definition, process, and challenges. In *Information Visualization*, volume 4950 of *Lecture Notes in Computer Science*, chapter 7, pages 154–175. Springer-Verlag, 2008.

[KAK95] Daniel A. Keim, Mihael Ankerst, and Hans-Peter Kriegel. Recursive pattern: A technique for visualizing very large amounts of data. In *VIS '95: Proceedings of the 6th conference on Visualization '95*, page 279, Washington, DC, USA, 1995. IEEE Computer Society.

[Kan00] Eser Kandogan. Star coordinates - a multi-dimensional visualization technique with uniform treatment of dimensions. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*. IEEE Computer Society, 2000.

[Kar53] Maurice Karnaugh. The map method for synthesis of combinational logic circuits. *Transactions of American Institute Of Electrical Engineers*, 72(9):593–599, 1953.

[KAS04] Daniel Keim, Mihael Ankerst, and Mike Sips. Visual data mining techniques. In Chris R. Johnson and Charles D. Hanson, editors, *Visualization Handbook*. Academic Press Inc., 2004.

[Kei00] Daniel A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6:59–78, 2000.

[Kei02] Daniel A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[KHG03] R. Kosara, H. Hauser, and D. Gresh. An interaktion view on information visualization. In *State-of-the-Art Proceedings of Eurographics*, 2003.

[KJ97] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

[KK06] Sotiris Kotsiantis and Dimitris Kanellopoulos. Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, 32(1):47–58, 2006.

[KKZ09] Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data*, 3:1:1–1:58, March 2009.

[KMG+06] Zoltán Konyha, Kresimir Matkovic, Denis Gracanin, Mario Jelovic, and Helwig Hauser. Interactive visual analysis of families of function graphs. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1373–1385, nov.-dec. 2006.

[KMH01] Robert Kosara, Silvia Miksch, and Helwig Hauser. Semantic depth of field. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*, San Diego, CA, USA, 22–23 2001.

[KMH09] Jörn Kohlhammer, Thorsten May, and Marcus Hoffmann. Visual analytics for the strategic decision making process. *Geospatial Visual Analytics - Geographic Information Processing and Visual Analyticsfor Environmental Security*, pages 299–310, 2009.

[KMS+08] Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. Visual analytics: Scope and challenges. In *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, pages 76–90. Springer-Verlag, Berlin, Heidelberg, 2008.

[KMSZ06] Daniel A. Keim, Florian Mansmann, Jorn Schneidewind, and Hartmut Ziegler. Challenges in visual data analysis. In *IV '06: Proceedings of the conference on Information Visualization*, pages 9–16, Washington, DC, USA, 2006. IEEE Computer Society.

[Koh95] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI, Montréal, Québec, Canada*, pages 1137–1145. Morgan Kaufmann Publishers, August 1995.

[Koh97] Teuvo Kohonen, editor. *Self-organizing maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1997.

[Kra93] Klaus-Peter Kratzer. *Neuronale Netze: Grundlagen und Anwendung*. Carl Hanser Verlag, München, Wien, 1993.

[Kum09] Ch. Aswani Kumar. Analysis of unsupervised dimensionality reduction techniques. *Comput. Sci. Inf. Syst.*, 6(2):217–227, 2009.

[KW78] Joseph B. Kruskal and Myron Wish. *Multidimensional Scaling*. Sage Publications, Inc., Newbury Park, CA, USA, 1978.

[Lam08] Heidi Lam. A framework of interaction costs in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1149–1156, 2008.

[Lan94] Pat Langley. Selection of relevant features in machine learning. In *In Proceedings of the AAAI Fall symposium on relevance*, pages 140–144. AAAI Press, 1994.

[LEW57] D. G. LEWIS. Normal distribution of intelligence: A critique. *British Journal of Psychology*, 48(2):98–104, 1957.

[LKL05] Jessica Lin, Eamonn Keogh, and Stefano Lonardi. Visualizing and discovering non-trivial patterns in large time series databases. *Information Visualization*, 4(2):61–82, 2005.

[LPWH06] John T. Langton, Astrid A. Prinz, David K. Wittenberg, and Timothy J. Hickey. Leveraging layout with dimensional stacking and pixelization to facilitate feature discovery and directed queries. In *Proceedings of the 1st Visual Information Expert Workshop (VIEW)*, volume 4370 of *Lecture Notes in Computer Science*, pages 77–91. Springer, 2006.

[LSG04] Danyu Liu, A.P. Sprague, and J.G. Gray. Polycluster: an interactive visualization approach to construct classification rules. In *Proceedings of the Third International Conference on Machine Learning and Applications (ICMLA)*, pages 280–287. IEEE Computer Society, 2004.

[LW06] J. Lehn and H. Wegmann. *Einführung in die Statistik*. Vieweg + Teubner, Wiesbaden, DE, 2006.

[Mac95] Alan MacEachren. *How Maps Work: Representation, Visualization, and Design*. The Guilford Press, New York, NY, 1995.

[Man97] Heikki Mannila. Methods and problems in data mining. In *ICDT '97: Proceedings of the 6th International Conference on Database Theory*, pages 41–55, London, UK, 1997. Springer-Verlag.

[May07] Thorsten May. Working with patterns in large multivariate datasets - karnaugh-veitch-maps revisited. In *IV '07: Proceedings of the 11th International Conference Information Visualization*, pages 277–285, Washington, DC, USA, 2007. IEEE Computer Society.

[MBD+11] Thorsten May, Andreas Bannach, James Davey, Tobias Ruppert, and Jörn Kohlhammer. Guiding feature subset selection with an interactive visualization *to appear*. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, Washington, DC, USA, 2011. IEEE Computer Society.

[MBN02] Luis Carlos Molina, Lluis Belanche, and Angela Nebot. Feature selection algorithms: a survey and experimental evaluation. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM)*, pages 306 – 313, Washington, DC, USA, 2002. IEEE Computer Society.

[McG95] J.E. McGrath. Methodology matters: Doing research in the behavioral and social sciences. In R. Baecker and W. A. S. Buxton, editors, *Readings in Human-Computer Interaction: An Interdisciplinary Approach. 2nd edition.*, pages 151–169. Morgan Kaufmann Publishers, San Mateo, CA, 1995.

[MD10] Thorsten May and Jörn Davey, James und Kohlhammer. Combining details of the chi-square goodness-of-fit test with multivariate data visualization. In *EuroVAST 2010, Proceedings of the 4th International Symposium on Advances in Visual Computing (ISVC)*, pages 45–50, Goslar, Germany, 2010. Eurographics Association.

[MDH+03] Alan M. MacEachren, Xiping Dai, Frank Hardisty, Diansheng Guo, and Eugene Lengerich. Exploring high-d spaces with multiform matrices and small multiples. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*. IEEE Computer Society, 2003.

[MDR11] Thorsten May, James Davey, and Tobias Ruppert. Smartstripes - looking under the hood of feature subset selection methods. In *EuroVAST 2011, Proceedings of the 4th International Symposium on Advances in Visual Computing (ISVC)*, pages 13–16, Goslar, Germany, 2011. Eurographics Association.

[Mir05] Boris Mirkin. *Clustering For Data Mining: A Data Recovery Approach*. Chapman & Hall, CRC Computer Science, 2005.

[MK08a] Thorsten May and Jörn Kohlhammer. Visual verification of hypotheses. In *Proceedings of the 4th International Symposium on Advances in Visual Computing (ISVC)*, volume 5359 of *Lecture Notes in Computer Science*, pages 31–42. Springer, 2008.

[MK08b] Thorsten May and Jörn Kohlhammer. Towards closing the analysis gap : Visual generation of decision supporting schemes from raw data. In *EuroVis '08: Computer Graphics Forum (Special Issue on Eurographics Symposium on Visualization)*, pages 277–285, Washington, DC, USA, 2008. IEEE Computer Society.

[MN06] Dharmesh M. Maniyar and Ian T. Nabney. Visual data mining using principled projection algorithms and information visualization techniques. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 643–648, New York, NY, USA, 2006. ACM.

[MNS06] Wolfgang Müller, Thomas Nocke, and Heidrun Schumann. Enhancing the visualization process with principal component analysis to support the exploration of trends. In *Proceedings of the Asia-Pacific Symposium on Information Visualisation (APVIS)*, pages 121–130, 2006.

[MRH+05] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(2):139–160, 2005.

[Mun09] Tamara Munzner. A nested process model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009.

[NHM+07] Eun Ju Nam, Yiping Han, Klaus Mueller, Alla Zelenyuk, and Dan Imre. Clustersculptor: A visual analytics tool for high-dimensional data. In *VAST '07: Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*, pages 75–82, Washington, DC, USA, 2007. IEEE Computer Society.

[Nor02] Donald A. Norman. *The Design of Everyday Things*. Basic Books, 2002. new print.

[NS98] Chris North and Ben Shneiderman. A taxonomy of multiple window coordinations. Technical report, University of Maryland, College Park, MD, 1998.

[NS00] Chris North and Ben Shneiderman. Snap-together visualization: a user interface for coordinating visualizations via relational schemata. In *AVI '00: Proceedings of the working conference on Advanced visual interfaces*, pages 128–135, New York, NY, USA, 2000. ACM.

[NSS05] Thomas Nocke, Stefan Schlechtweg, and Heidrun Schumann. Icon-based visualization using mosaic metaphors. In *Proceedings of the Ninth International Conference on Information Visualisation (IV)*, pages 103–109, Washington, DC, USA, 2005. IEEE Computer Society.

[OD08] David L. Olson and Dursun Delene. Fuzzy sets in data mining. In *Advanced Data Mining Techniques*, pages 69–86. Springer, Berlin, Heidelberg, 2008.

[PB80] Jr. W. D. Perreault and Jr. H. C. Barksdale. A model-free approach for analysis of complex contingency data in survey research. *Journal of Marketing Research*, 17:503–515, 1980.

[PBH08] Harald Piringer, Wolfgang Berger, and Helwig Hauser. Quantifying and comparing features in high-dimensional datasets. In *Proceedings of the 2008 12th International Conference Information Visualisation*, pages 240–245, Washington, DC, USA, 2008. IEEE Computer Society.

[PBK10] H. Piringer, W. Berger, and J. Krasser. Hypermoval: Interactive visual validation of regression models for real-time simulation. *Computer Graphics Forum*, 29(3), 2010.

[Pla04] Catherine Plaisant. The challenge of information visualization evaluation. In *AVI '04: Proceedings of the working conference on Advanced visual interfaces*, pages 109–116, New York, NY, USA, 2004. ACM.

[PSCO09] W. A. Pike, J. Stasko, R. Chang, and T. O'Connell. The science of interaction. *Information Visualization Special Issue: Foundations and Frontiers of Visual Analytics*, 8(4):263–274, 2009.

[PZL+05] Yanwei Pang, Lei Zhang, Zhengkai Liu, Nenghai Yu, and Houqiang Li. Neighborhood preserving projections (NPP): A novel linear dimension reduction method. In De-Shuang Huang, Xiao-Ping Zhang, and Guang-Bin Huang, editors, *Proceedings of the International Conference on Intelligent Computing, ICIC, Hefei, China, August 23-26*, Lecture Notes in Computer Science, pages 117–125. Springer, 2005.

[Qui87] J. R. Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3):221–234, 1987.

[Qui96] J. R. Quinlan. Bagging, boosting, and c4.5. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 725–730. AAAI Press, 1996.

[RC94] Ramana Rao and Stuart K. Card. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 318–322, New York, NY, USA, 1994. ACM.

[RK04] M. Rasmussen and G. Karypis. gcluto - an interactive clustering, visualization and analysis system. Technical report, University of Minnesota, MN, US, 2004. Technical Report 04-021.

[RMKS10] Tobias Ruppert, Thorsten May, Jörn Kohlhammer, and Tobias Schreck. Visuelle analysen des datensatzes: Wie versteckte zusammenhänge sichtbar werden. *Allgemeinbildung in Deutschland - Erkenntnisse aus dem Spiegel-StudentenPISA Test*, pages 87–104, 2010.

[RTT05] José Fernando Jr. Rodrigues, Agma J. M. Traina, and Caetano Jr. Traina. Visualization tree, multiple linked analytical decisions. In *Smart Graphics*, pages 65–76, 2005.

[SBvLK09] Tobias Schreck, Jürgen Bernard, Tatiana von Landesberger, and Jörn Kohlhammer. Visual cluster analysis of trajectory data with interactive kohonen maps. *Information Visualization*, 8(1):14–29, 2009.

[Sch93] Cullen Schaffer. Overfitting avoidance as bias. *Mach. Learn.*, 10(2):153–178, 1993.

[Sch07] Jörn Schneidewind. *Scalable Visual Analytics : Solutions and Techniques for Business Applications*. PhD thesis, Universität Konstanz, Universitätsstr. 10, 78457 Konstanz, 2007.

[SHF96] Eberhard Schöneburg, Frank Heinzmann, and Sven Feddersen. *Genetische Algorithmen und Evolutionsstrategien*. Addison-Wesley, Bonn, Paris, Reading, Mass., 1996.

[Shn92] Ben Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph.*, 11(1):92–99, 1992.

[Shn96] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *VL '96: Proceedings of the 1996 IEEE Symposium on Visual Languages*, pages 336–343, Washington, DC, USA, 1996. IEEE Computer Society.

[Shn02] Ben Shneiderman. Inventing discovery tools: combining information visualization with data mining. *Information Visualization*, 1(1):5–12, 2002.

[Sii00] Harri Siirtola. Direct manipulation of parallel coordinates. In *CHI '00: CHI '00 extended abstracts on Human factors in computing systems*, pages 119–120, New York, NY, USA, 2000. ACM.

[Sir06] M. Sirkin. *Statistics for the Social Sciences, 3rd Edition*. Sage Publication Inc., Thousand Oaks, CA, 2006.

[SLV04] Vojtech Svátek, Martin Labský, and Miroslav Vacura. Knowledge modelling for deductive web mining. In *Proceedings of the 14th Engineering Knowledge in the Age of the Semantic Web, EKAW 2004, Whittlebury Hall, UK, October 5-8*, volume 3257 of *Lecture Notes in Computer Science*, pages 337–353. Springer, 2004.

[SM99] Nigel Shadbolt and Nick Milton. From knowledge engineering to knowledge management. *British Journal of Management*, 10:309–322, 1999.

[SM05] Harri Siirtola and Erkki Mäkinen. Constructing and reconstructing the reorderable matrix. *Information Visualization*, 4(1):32–48, 2005.

[SP04] Ben Shneiderman and Catherine Plaisant. *Designing the User Interface: Strategies for Effective Human-Computer Interaction (4th Edition)*. Pearson Addison Wesley, 2004.

[SR95] R.R. Sokal and F.J. Rohlf. *Biometry: The principles and practice of statistics in biological research, 3rd Edition*. W.H. Freeman, New York, 1995.

[SS04a] Jinwook Seo and Ben Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pages 65 –72, 0-0 2004.

[SS04b] Steffen Staab and Rudi Studer, editors. *Handbook on Ontologies*. International Handbooks on Information Systems. Springer, 2004.

[SS07] Hideaki Shimazaki and Shigeru Shinomoto. A method for selecting the bin size of a time histogram. *Neural Comput.*, 19(6):1503–1527, June 2007.

[SSKS06] Mike Sips, Jörn Schneidewind, Daniel A. Keim, and Heidrun Schumann. Scalable pixel-based visual interfaces: Challenges and solutions. In *Proceedings of the 10th International Conference on Information Visualisation, (IV)*, pages 32–38. IEEE Computer Society, 2006.

[SSM05] Bartlomiej Sniezynski, Robert Szymacha, and Ryszard S. Michalski. Knowledge visualization using optimized general logic diagrams. In *Proceedings of the International Conference on Intelligent Information Processing and Web Mining (IIPWM)*, Advances in Soft Computing, pages 137–146. Springer, 2005.

[STH02] Chris Stolte, Diane Tang, and Pat Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):52–65, 2002.

[SvW08] Yedendra Babu Shrinivasan and Jarke J. van Wijk. Supporting the analytical reasoning in information visualization. In *Proceedings of the 26th annual SIGCHI conference on Human Factors in computing systems*, CHI'08, pages 1237–1246, New York, NY, USA, 2008. ACM.

[TC05] James J. Thomas and Kristin A. Cook. *Illuminating the Path*. IEEE CS Press, 2005.

[TG80] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, January 1980.

[THM+05] Judi Thomson, Elizabeth Hetzler, Alan MacEachren, Mark Gahegan, and Misha Pavel. A typology for visualizing uncertainty. *Visualization and Data Analysis 2005*, 5669(1):146–157, 2005.

[TKB07] Ulanbek D. Turdukulov, Menno-Jan Kraak, and Connie A. Blok. Designing a visual environment for exploration of time series of remote sensing data: In search for convective clouds. *Computers and Graphics*, 31:410–428, 2007.

[TM03] Soon Tee Teoh and Kwan-Liu Ma. Starclass: Interactive visual classification using star coordinates. In *Proceedings of the Third SIAM International Conference on Data Mining*. SIAM, 2003.

[TSDS96] Lisa Tweedie, Robert Spence, Huw Dawkes, and Hua Su. Externalising abstract mathematical models. In *CHI*, pages 406–, 1996.

[Tuf83] Edward Tufte. *The Visual Display of Quantitive Information.* Graphics Press, Cheshire, Conn., 1983.

[Tuk77] J. W. Tukey. *Exploratory Data Analysis.* Addison-Wesley, Bonn, Paris, Reading MA, 1977.

[VPF06] Eliane R. A. Valiati, Marcelo S. Pimenta, and Carla M. D. S. Freitas. A taxonomy of tasks for guiding the evaluation of multidimensional visualizations. In *BELIV '06: Proceedings of the 2006 AVI workshop on BEyond time and errors*, pages 1–6, New York, NY, USA, 2006. ACM.

[VT07] Lucian Voinea and Alex Telea. Visual data mining and analysis of software repositories. *Computers and Graphics*, 31:410–428, 2007.

[vW05] Jarke J. van Wijk. The value of visualization. In *16th IEEE Visualization Conference (VIS 2005), 23-28 October 2005, Minneapolis, MN, USA*, page 11. IEEE Computer Society, 2005.

[VWvdW99] Jarke J. Van Wijk and Huub van de Wetering. Cushion treemaps: Visualization of hierarchical information. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*, page 73, Washington, DC, USA, 1999. IEEE Computer Society.

[WAG05] Leland Wilkinson, Anushka Anand, and Robert Grossman. Graph-theoretic scagnostics. In *INFOVIS '05: Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, page 21, Washington, DC, USA, 2005. IEEE Computer Society.

[WAG06] Leland Wilkinson, Anushka Anand, and Robert Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1363–1372, 2006.

[War04a] Matthew Ward. Finding needles in large-scale multivariate data haystacks. *IEEE Comput. Graph. Appl.*, 24(5):16–19, 2004.

[War04b] Colin Ware. *Information Visualization: Perception for Design.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004.

[WB98] Christopher Westphal and Teresa Blaxton. *Data mining solutions: methods and tools for solving real-world problems.* John Wiley & Sons, Inc., New York, NY, USA, 1998.

[WF09] Leland Wilkinson and Michael Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.

[WFH+01] Malcolm Ware, Eibe Frank, Geoffrey Holmes, Mark Hall, and Ian H. Witten. Interactive machine learning: letting users build classifiers. *Int. J. Hum.-Comput. Stud.*, 55(3):281–292, 2001.

[WLT94] M.O. Ward, J. LeBlanc, and R. Tipnis. N-land : A graphical tool for exploring n-dimensional data. In *Proceeding of the Computer Graphics International Conference*, 1994.

[WM94] J. Wnek and R. Michalski. Comparing symbolic and subsymbolic learning: three studies. In R. Michalski and G. Tecuci, editors, *Machine Learning: A Multistrategy Approach*, volume 4. Morgan-Kaufman, 1994.

[WOWR03] Jörg Walter, Jörg Ontrup, Daniel Wessling, and Helge Ritter. Interactive visualization and navigation in large data collections using the hyperbolic space. In *ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining*, page 355, Washington, DC, USA, 2003. IEEE Computer Society.

[WSB92] B. J. Wielinga, A. Th. Schreiber, and J. A. Breuker. KADS: a modelling approach to knowledge engineering. *Knowl. Acquis.*, 4(1):5–53, 1992.

[Yao03] Yiyu Y. Yao. A step toward the foundations of data mining. *Data Mining and Knowledge Discovery: Theory, Tools, and Technology V*, 5098(1):254–263, 2003.

[YH98] Jihoon Yang and Vasant G. Honavar. Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems*, 13(2):44–49, 1998.

[YHW+07] Jing Yang, Daniel Hubball, Matthew O. Ward, Elke A. Rundensteiner, and William Ribarsky. Value and relation display: Interactive visual exploration of large data sets with hundreds of dimensions. *IEEE Transactions on Visualization and Computer Graphics*, 13:494–507, 2007.

[YKSJ07] Ji Soo Yi, Youn ah Kang, John Stasko, and Julie Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, 2007.

[YWRH03] J. Yang, M. O. Ward, E. A. Rundensteiner, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets. In *Proceedings of the symposium on Data visualisation (VisSym)*, pages 19–28, Aire-la-Ville, Switzerland, Switzerland, 2003. Eurographics Association.

[YXRW07] Di Yang, Zaixian Xie, Elke A. Rundensteiner, and Matthew O. Ward. Managing discoveries in the visual analytics process. *SIGKDD Explor. Newsl.*, 9(2):22–29, 2007.

[ZC07] Torre Zuk and Sheelagh Carpendale. Visualization of uncertainty and reasoning. In *SG '07: Proceedings of the 8th international symposium on Smart Graphics*, pages 164–177, Berlin, Heidelberg, 2007. Springer-Verlag.

Anhang A: Lebenslauf

Name:	Thorsten May	
Geburtstag:	5. Januar 1975	
Geburtsort:	Offenbach	
Schulbildung:	1981 – 1985	Grundschule, Rote-Warthe-Schule/Mühlheim
	1985 – 1987	Förderstufe, Goetheschule/Mühlheim
	1987 – 1994	Gymnasium, Friedrich-Ebert-Gymnasium/Mühlheim, (Abschluss Abitur)
Studium:	Oktober 1994 – Oktober 2000	Mathematik mit Schwerpunkt Informatik, an der Technischen Universität Darmstadt (TUD), (Abschluss Diplom-Mathematik)
Berufspraxis:	1. März 2001 – 1. September 2001	Wissenschaftliche Hilfskraft mit Abschluss, Fraunhofer IGD, Abteilung Animation & Bildkommunikation (A3), Darmstadt
	seit 1. Dezember 2001	Wissenschaftlicher Mitarbeiter, Fraunhofer IGD, Abteilung Animation & Bildkommunikation (A3), Darmstadt
	seit 1. April 2008	Stellvertretender Abteilungsleiter, Fraunhofer IGD, Abteilung Informationsvisualisierung & Visual Analytics (IVA), Darmstadt